

---

# **KEPLER**

# **DATA PROCESSING**

# **HANDBOOK**

## **KSCI-19081-002**

---

**Edited by Jon M. Jenkins**  
NASA Ames Research Center



NASA Ames Research Center  
Moffett Field CA 94035



Edited by:



1/20/17

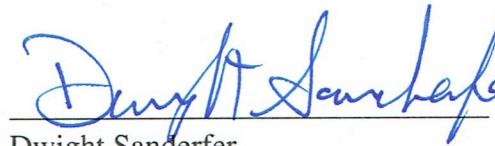
Date

Jon M. Jenkins

Co-I for Data Processing

NASA Ames Research Center

Approved by:

  
Dwight Sanderfer

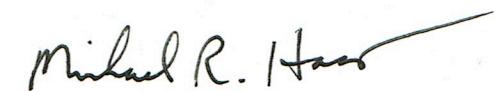
1/20/17

Date

Science Operations Center Manager

NASA Ames Research Center

Approved by:

  
Michael R. Haas

1/20/17

Date

Science Office Director

NASA Ames Research Center

Approved by:

  
Natalie M. Batalha

25 Jan 2017

Date

Project Scientist

NASA Ames Research Center



# **Document Control**

## **Ownership**

This document is part of the Kepler Project Documentation that is controlled by the Kepler Project Office, NASA/Ames Research Center, Moffett Field, California.

## **Control Level**

This document will be controlled under KPO @ Ames Configuration Management system. Changes to this document shall be controlled.

## **Physical Location**

The physical location of this document will be in the KPO @ Ames Data Center.

## **Distribution Requests**

To be placed on the distribution list for additional revisions of this document, please address your request to the *Kepler* Science Operations Center:

Jon M. Jenkins  
Kepler Co-Investigator for Data Analysis  
MS 269-3  
NASA Ames Research Center  
Moffett Field, CA 94035-1000  
Jon.M.Jenkins@nasa.gov

When citing this document as a whole, please use the following reference:

Jenkins, J. M., (ed.) 2017. *Kepler* Data Processing Handbook: KSCI-19081-002

Each chapter in this handbook has a separate author list and is designed to stand alone. Consequently, please reference the relevant chapters with a format similar to this example for Chapter 4 on Dynablack:

Clarke, B. D., et al. 2017. “Dynamic Black Correction,” in *Kepler* Data Processing Handbook: KSCI-19081-002, Jenkins, J. M. (ed.), 81–105



*To David G. Koch*



# CONTENTS

---

Acknowledgements	xv
Contributors	xvii
Preface	xix

## PART I KEPLER SCIENCE OPERATIONS

<b>1 Philosophy and Scope</b>	<b>3</b>
1.1 Intended Audience	3
1.2 Relationship to Other Documents	4
1.2.1 <i>Kepler</i> Instrument Handbook (KSCI-19033)	4
1.2.2 <i>Kepler</i> Data Characteristics Handbook (KSCI-19040)	4
1.2.3 <i>Kepler</i> Data Release Notes (KSCI-19041 – KSCI-19065)	4
1.2.4 <i>Kepler</i> Archive Manual (KDMC-10008)	5
1.2.5 <i>Kepler</i> Input Catalog	5
1.3 Document Organization – From Pixels to Planets	5
Bibliography	7
<b>2 Overview of the Science Operations Center</b>	<b>9</b>
2.1 Introduction	9
2.2 The <i>Kepler</i> Mission Ground Segment	13
2.3 The Science Operations Center	14
2.3.1 Software Infrastructure	15
2.3.2 Photometer Management	17
2.3.3 Science Pipeline	18
2.3.4 Commissioning Tools	21
2.3.5 Hardware	22
2.3.6 Cluster Datastore Machine	22
2.4 Running the Pipeline	23
2.4.1 Unit of Work	23
2.4.2 Coordination of Work	24
2.4.3 Remote Execution	25
2.4.4 Triggers	26
2.4.5 Data Accountability	26

2.4.6	Operating the Pipeline	26
2.5	Conclusions	28
	Bibliography	30
<b>3</b>	<b>TAD: Selecting Pixels for Downlink</b>	<b>35</b>
3.1	Introduction	35
3.1.1	The <i>Kepler</i> Focal Plane	36
3.1.2	Target and Aperture Definitions Task Flow	37
3.1.3	<i>Kepler</i> Pixel and Target Types	38
3.1.4	Pixel Selection Requirements	39
3.2	Pixel Selection	40
3.2.1	Synthetic Image Creation	40
3.2.2	Optimal Pixel Selection	44
3.2.3	Background Pixel Selection	45
3.3	Mask Creation and Assignment	46
3.3.1	Required Pixels: Adding Pixel Margin	47
3.3.2	Mask Table Creation	48
3.3.3	Mask Assignment	49
3.4	Reference Pixel Targets	49
3.5	Conclusions	51
	Bibliography	51
<b>PART II THE KEPLER PHOTOMETRIC PIPELINE</b>		
<b>4</b>	<b>Dynamic Black Correction</b>	<b>57</b>
4.1	Introduction	57
4.2	Background	58
4.2.1	Description of Image Artifacts	59
4.2.2	<i>Kepler</i> 's Noise Floor	61
4.2.3	Acceptable Bias Variations	63
4.2.4	Artifact Removal and Flagging Objectives	63
4.3	Methods	64
4.3.1	Spatial Fitting	64
4.3.2	Thermo-Temporal Fitting	69
4.3.3	2-D Black Correction	69
4.3.4	RBA Flagging	71
4.4	Results	72
4.4.1	Example Fits	72
4.4.2	Effect of Corrections on Targets	73
4.4.3	Flagging Effectiveness	76
	Appendix A: Terms in the Spatial Model	77
	Bibliography	79
<b>5</b>	<b>Pixel Level Calibrations</b>	<b>81</b>

5.1	Introduction	81
5.2	<i>Kepler</i> Data Formats and CAL Unit of Work	82
5.2.1	CAL Data Types: Long and Short Cadence and Full Frame Images	82
5.2.2	Focal Plane Array	83
5.2.3	Pixel Collection	83
5.2.4	Photometric and Collateral Data	84
5.2.5	Processing Order	85
5.2.6	Data Gaps	87
5.3	Calibration	89
5.3.1	Compute Raw Black Uncertainties	92
5.3.2	Models	93
5.3.3	Fixed Offset, Mean Black, and Spatial Co-Adds	94
5.3.4	Black Correction	95
5.3.5	Nonlinearity and Gain Correction	98
5.3.6	LDE Overshoot/Undershoot Correction	99
5.3.7	Smear and Dark Correction	100
5.3.8	Flat Field Correction	102
5.3.9	Additional Functionality in CAL	103
5.4	Summary	103
	Bibliography	104
<b>6</b>	<b>Photometric Analysis</b>	<b>107</b>
6.1	Introduction	107
6.2	Architecture	109
6.3	Photometric Analysis Science Algorithms	112
6.3.1	Cosmic Ray Cleaning	113
6.3.2	Background Estimation and Removal	118
6.3.3	Centroding	119
6.3.4	Motion Polynomial Fitting	121
6.3.5	Optimal Aperture Selection	123
6.3.6	Simple Aperture Photometry	127
6.4	Summary and Conclusions	129
	Appendix A: Obtaining Centroid and Plate Scale Estimates from Motion Polynomials	130
	Bibliography	130
<b>7</b>	<b>Finding Optimal Apertures in Kepler Data</b>	<b>133</b>
7.1	Introduction	133
7.2	Finding Optimal Apertures	134
7.3	Method #1: Using the Synthetic FFI and a Pure PRF Image Model	135
7.4	Method #2: Using a PRF-Based Image Model and the Pixel Scene	138
7.4.1	Modeling Target Masks	140
7.4.2	Optimizing the Aperture for Photometric Precision	142
7.5	Selecting the Best Aperture using CDPP and Logistic Regression	144

7.6	Per Cadence Flux Fraction and Crowding Metric	146
7.7	Saturated Pixels	146
7.8	Summary and Example of Performance	147
7.9	Application to K2 Data	151
7.10	Conclusions	155
	Bibliography	157
<b>8</b>	<b>Presearch Data Conditioning</b>	<b>159</b>
8.1	The <i>Kepler</i> SOC Pre-Search Data Conditioning Pipeline Module	159
8.2	Introduction	160
8.2.1	Errors in the Light Curves: Tasks of PDC	161
8.3	Architecture and Algorithms of PDC	162
8.3.1	Cotrending of Systematic Errors in PDC	162
8.3.2	Inputs to PDC	168
8.3.3	Overview and Data Flow	170
8.3.4	Outputs of PDC	184
8.3.5	Processing Times	185
8.4	A Bayesian Approach to Correcting Systematic Errors	186
8.4.1	The Basic Problem and the Principle Behind the Solution	186
8.4.2	The MAP Approach, An Analytical Solution	190
8.5	The Empirical Bayesian MAP Approach and Implementation	192
8.5.1	Finding the Cotrending Basis Vectors	193
8.5.2	Numerically Generating $p(\theta)$	199
8.5.3	Finding the Weighting Parameter $\mathbf{W}_{\text{pr}}$	204
8.5.4	Maximization of the Posterior PDF	206
8.5.5	Iterating the Posterior with the Goodness Metric	208
8.5.6	Propagation of Uncertainties	210
8.5.7	Application of PDC-MAP to Short Cadence Data	211
8.6	Multi-Scale MAP	212
8.6.1	Shortcomings of PDC-MAP	213
8.6.2	The Solution: Multiscale Error Correction	216
8.6.3	Choice of Parameters	221
8.6.4	Further Algorithm Details	227
8.6.5	Performance Evaluation	231
8.6.6	Outlook	233
8.7	Conclusions	234
	Bibliography	235

### PART III TRANSIT SEARCH ENGINE

<b>9</b>	<b>Transiting Planet Search</b>	<b>241</b>
9.1	Introduction	241
9.2	Light Curve Preprocessing	245

9.2.1	Identification of Outliers and Astrophysical Transients	245
9.2.2	Identifying and Removing Phase-Shifting Harmonics	245
9.2.3	Edge Detrending of Contiguous Blocks of Flight Data	247
9.3	Generation of Single Event Statistics and CDPP	248
9.3.1	A Wavelet-Based Matched Filter	248
9.3.2	Removal of Positive Flux Outliers	251
9.3.3	Quarter-by-Quarter Whitening	254
9.3.4	Setting De-emphasis Weights	254
9.4	Folding the Detection Statistics and Applying Veto	254
9.4.1	Folding the Single Event Statistics	255
9.4.2	Search Templates and Template Spacing	256
9.4.3	Limitation on Allowable Transit Duty Cycles	256
9.4.4	False Alarm Veto	258
9.4.5	Removal of Non-Periodic Transit-Like Features	261
9.5	Performance of TPS in the DR25 Search	262
9.5.1	Incomplete Searches	262
9.5.2	Detection of Multiple-Planet Systems	263
9.5.3	TCE Population	264
9.5.4	Comparison with Known Kepler Objects of Interest (KOIs)	268
9.5.5	Matching of Golden KOI and TCE Ephemerides	271
9.6	Conclusions	275
Appendix A: Generation of the Non-Decimated Octave-Band Filters		275
Appendix B: Determining the Threshold for Positive Outlier Removal		277
Bibliography		277
<b>10</b>	<b>A Statistical Bootstrap Test</b>	<b>281</b>
10.1	Introduction	281
10.2	Detection Algorithm	283
10.3	Bootstrap Test	287
10.4	Archive Column Definitions	289
10.5	Results	290
10.5.1	Q1–Q16	290
10.5.2	SOC 9.2 Q1–Q17 DR24	291
10.5.3	SOC 9.3 Q1–Q17 DR25	295
10.5.4	Comparison Across Datasets	296
10.6	Precision of the Statistical Bootstrap Results	299
10.7	Bootstrap Analysis of a Single TCE	303
10.8	Conclusions	303
Bibliography		304
<b>11</b>	<b>Data Validation I – Diagnostic Tests</b>	<b>307</b>
11.1	Introduction	307
11.1.1	Vetting Transiting Planet Candidates	308

11.2	Pipeline Data Validation	309
11.3	Diagnostic Tests	311
11.3.1	Weak Secondary Test	312
11.3.2	Rolling Band Diagnostic	313
11.3.3	Eclipsing Binary Discrimination Tests	314
11.3.4	Difference Imaging and Centroid Offset Analysis	316
11.3.5	Statistical Bootstrap	321
11.3.6	Centroid Motion Test	322
11.3.7	Optical Ghost Diagnostic Test	325
11.4	KOI Matching	325
11.5	Archive Products	325
11.5.1	DV Report	326
11.5.2	DV Report Summary	329
11.5.3	DV Time Series	330
11.6	Conclusion	330
	Bibliography	330
<b>12</b>	<b>Data Validation II – Transit Model Fitting</b>	<b>333</b>
12.1	Introduction	333
12.2	Light Curve Preprocessing	336
12.2.1	Baseline Removal and Light Curve Normalization	336
12.2.2	Quarterly Data Segment Stitching	336
12.2.3	Harmonic Removal	337
12.2.4	Timestamp Conversion	337
12.3	Geometric Transit Model	338
12.3.1	Fitted Parameters	338
12.3.2	Derived Parameters	339
12.4	Geometric Transit Signal Generator	340
12.5	Geometric Model Fitting Algorithms	343
12.5.1	Iterative Whitening and Model Fitting	343
12.5.2	Reduced Parameter Fits	352
12.5.3	Odd-Even Transit Fit	353
12.5.4	Outputs of Geometric Transit Model Fits	356
12.5.5	Alerts of Failed Fitting Cases	358
12.6	Trapezoidal Model Fitting Algorithm	358
12.7	Multiple-Planet Search	360
12.8	Performance of Transit Model Fitting	360
12.9	Conclusions	364
	Bibliography	364
<b>13</b>	<b>Acronyms and Abbreviations</b>	<b>367</b>

# ACKNOWLEDGEMENTS

---

Many individuals have contributed to the *Kepler* Science Data Processing Pipeline over the last 13 years. We note these individuals for their contributions.

## **Science Operations Center**

Jon M. Jenkins (Science Lead), Christopher K. Middour (Systems Engineer), David P. Pletcher (Manager), Dwight Sanderfer (Manager),

## **SOC Scientific Programmers**

Joseph D. Twicken (Lead), Peter Tenenbaum, Bruce D. Clarke, Jie Li, Robert L. Morris, Shawn Seader, Jeffrey C. Smith, Joseph H. Catanzarite, Elisa V. Quintana, Hayley Wu, Hema Chandrasekaran

## **SOC Software Developers**

Todd C. Klaus (Lead), Sean D. McCauliff (Lead), Miles T. Cote, Forrest Girouard, Bill Wohler, Christopher L. Allen, Lee S. Brownston, Jay P. Gunter

## **SOC Operations Staff**

Jennifer Campbell (Lead), Khadeejah Zamudio, AKM Kamal Uddin, Brett Stroozas (Lead Ops Engineer)

## **Science Office**

Michael R. Haas (Director), Jessie L. Dotson (Deputy Director), Stephen T. Bryson, Christopher J. Burke, Douglas A. Caldwell (Instrument Scientist), Jeffrey L. Coughlin, Jessie L. Christiansen, Jeff Kolodziejczak (MSFC), Pavel Machalak, Fergal Mullally, Jason F. Rowe, Susan E. Thompson, Jeffrey Van Cleve

## **Guest Observer Office**

Thomas Barclay (Director), Michael N. Fanelli (Deputy Director), Karen Kinemuchi, Martin D. Still (Director)

## **Scientists At Large**

Ronald L. Gilliland (STSci – Science Team)  
Natalie M. Batalha (Mission Scientist)  
David G. Koch (Deputy Science PI)  
William J. Borucki (Science PI)

We offer many thanks to our technical writer, Veronica Phillips for her help in reviewing and editing this handbook. We also thank the greater *Kepler* team for their contributions.

# CONTRIBUTORS

---

## **Chapter 1 – Philosophy and Scope**

Jon M. Jenkins

## **Chapter 2 – Overview of the Science Operations Center**

Jon M. Jenkins

## **Chapter 3 – Target and Aperture Definitions: Selecting Pixels for Kepler Downlink**

Stephen T. Bryson, Jon M. Jenkins, Todd C. Klaus, Miles T. Cote, Elisa V. Quintana, Jennifer R. Campbell, Khadeejah Zamudio, Hema Chandrasekaran, Douglas A. Caldwell, Jeffrey E. Van Cleve, and Michael R. Haas

## **Chapter 4 – Dynamic Black Correction**

Bruce D. Clarke, Jeffrey J. Kolodziejczak, Douglas A. Caldwell, Jeffrey E. Van Cleve, Jon M. Jenkins, Miles T. Cote, Todd C. Klaus, and Vic S. Argabright

## **Chapter 5 – Pixel Level Calibrations**

Bruce D. Clarke, Douglas A. Caldwell, Elisa V. Quintana, Hema Chandrasekaran, Joseph D. Twicken, Jon M. Jenkins, Miles T. Cote, Sean D. McCauliff, Todd C. Klaus, Christopher Allen, and Stephen T. Bryson

## **Chapter 6 – Photometric Analysis: Algorithms and Architecture**

Robert L. Morris, Joseph D. Twicken, Jeffrey C. Smith, Bruce D. Clarke, Jon M. Jenkins, Stephen T. Bryson, Forrest Girouard, and Todd C. Klaus

## **Chapter 7 – Finding Optimal Apertures in Kepler Data**

Jeffrey C. Smith, Robert L. Morris, Jon M. Jenkins, Stephen T. Bryson, Douglas A. Caldwell, and Forrest R. Girouard

## **Chapter 8 – Presearch Data Conditioning**

Jeffrey C. Smith, Martin C. Stumpe, Jon M. Jenkins, Jeffrey E. Van Cleve, Forrest R. Girouard, Jeffery J. Kolodziejczak, Sean D. McCauliff, Robert L. Morris, and Joseph D. Twicken

## **Chapter 9 – Transiting Planet Search**

Jon M. Jenkins, Peter Tenenbaum, Shawn Seader, Christopher J. Burke, Sean D. McCauliff, Jeffrey C. Smith, Joseph D. Twicken, and Hema Chandrasekaran

**Chapter 10 – A Computationally Efficient Statistical Bootstrap Test for Transiting Planets**

Jon M. Jenkins, Shawn Seader, and Christopher J. Burke

**Chapter 11 – Data Validation I – Diagnostic Tests**

TBD

**Chapter 12 – Data Validation II – Transit Model Fitting and Multiple Planet Search**

Jie Li, Peter Tenenbaum, Joseph D. Twicken, Christopher J. Burke, Jon M. Jenkins, Elisa V. Quintana, Jason F. Rowe, and Shawn E. Seader

# PREFACE

---

The *Kepler* Mission collected four years of data on 208,000 stars from May 13, 2009 until May 11, 2013. The first edition of the *Kepler* Data Processing Handbook collected together a set of SPIE papers and other manuscripts documenting the SOC 6.2 codebase, which was the first complete version of the *Kepler* Science Operations Center Pipeline. At that time, the SOC pipeline did not search light curves for multiple planets (that capability didn't arrive until SOC 8.2 in 2012).

*Kepler* stressed the technologies needed to collect, process and analyze the data and to conduct follow up observations, pushing the envelope on all fronts. Scientists involved in every step of the discovery process reported that they had to overhaul and extend their data analysis tools to deal effectively with the *Kepler* data and the intriguing *Kepler* objects. Needless to say, the *Kepler* science data processing pipeline underwent significant evolution both during and after the data collection phase. This volume represents the most current information about the as-built SOC 9.3 algorithms used to generate *Kepler*'s final legacy archive data products.

We note that Chapter 11, “Data Validation I Architecture, Diagnostic Tests, and Data Products,” is a preliminary version that will be completed and included in the third edition of the *Kepler* Data Processing Handbook as soon as it is available.

JON M. JENKINS

*Moffett Field, CA U.S.A*

*January 18, 2017*



## PART I

---

# OVERVIEW OF KEPLER SCIENCE OPERATIONS



# CHAPTER 1

---

## PHILOSOPHY AND SCOPE

JON M. JENKINS

NASA Ames Research Center, Moffett Field, CA 94035

### 1.1 Intended Audience

The *Kepler* Data Processing Handbook (KDPH) is written for a broad audience: the astrophysics and exoplanet science communities, archival researchers, peer reviewers of *Kepler* results submitted for publication, and also for engineers and scientists who are designing their own science data processing pipelines for wide-angle transit surveys, such as the upcoming TESS and PLATO missions. The KDPH presents an overview of the *Kepler* Science Operations Center (SOC), describing the architecture and functionality supported by this facility and software codebase. The KDPH then describes the selection of pixels for storage and downlink from the *Kepler* spacecraft for the up to 170,000 target stars that could be acquired at any time on *Kepler*. These target pixel datasets are collected by the Flight System as described in the *Kepler* Instrument Handbook (KIH – Van Cleve & Caldwell, 2016). The KDPH then describes the transformation of these pixel sets into photometric time series and the detection and validation of transits in those time series by the science data processing pipeline. The KDPH provides information about the algorithms, inputs, outputs, and performance of target management and pipeline software components, called Computer Software Configuration Items (CSCIs).

The primary goal of the KDPH is to document the algorithms used to produce the *Kepler* legacy archive calibrated pixel and light curve data products at the Mikulski Archive for Space Telescopes (MAST)<sup>1</sup> and the Data Validation (DV) reports and diagnostic information for potential planet candidates, called Threshold Crossing Events (TCEs), archived at the NASA Exoplanet Archive

<sup>1</sup>Available at <https://archive.stsci.edu/kepler/>.

(Akeson et al., 2013).<sup>2</sup> Additional *Kepler* Mission documentation of interest for interpreting and analyzing the *Kepler* data products includes the KIH, the *Kepler* Archive Manual (KAM – Thompson et al., 2016), the *Kepler* Data Characteristics Handbook (KDCH – Van Cleve et al., 2016), and associated Data Release Notes.<sup>3</sup> These and other documentation of potential interest can be found at <https://archive.stsci.edu/kepler/documents.html>. *Kepler* and K2 Mission news and additional documentation can be found at <https://keplerscience.arc.nasa.gov>, which also provides information and tools for observation planning and data analysis. In addition, the *Kepler* SOC codebase will be available as a NASA Open Source project prior to September 30, 2017 for those researchers who wish to understand the algorithms and functionality at the deepest levels by reading the source files, tracing the code and/or implementing the science algorithms used to manage and produce the *Kepler* archive data products, conduct target and photometer management, and/or design commissioning tools. Together, the KDPH, the SOC codebase, and the *Kepler* documents listed above supply the information necessary for understanding and interpreting *Kepler* results, given the real properties of the hardware and the as-built data analysis methods.

## 1.2 Relationship to Other Documents

### 1.2.1 *Kepler* Instrument Handbook (KSCI-19033)

The KIH (Van Cleve & Caldwell, 2016) provides information about the design, performance, and operational constraints of the *Kepler* hardware as well as an overview of the available pixel datasets. It presents an overview of the *Kepler* instrument, and then tracks photons through the telescope, focal plane, and focal plane electronics. Details regarding targets, the pixels of interest around them, and operational details are specified, which will be helpful in both planning observations for the repurposed *Kepler* Mission, dubbed K2, and for understanding the data reduction procedures described later in this document.

### 1.2.2 *Kepler* Data Characteristics Handbook (KSCI-19040)

The Data Characteristics Handbook (Van Cleve et al., 2016) provides a description of a variety of phenomena identified within the *Kepler* data and a discussion of how these phenomena are handled by the science data processing pipeline.

### 1.2.3 *Kepler* Data Release Notes (KSCI-19041 – KSCI-19065)

With each release of data, a set of accompanying notes was created to give *Kepler* users information specific to the time period during which the data was ob-

<sup>2</sup> Available at <http://exoplanetarchive.ipac.caltech.edu/>.

<sup>3</sup> Available at [https://archive.stsci.edu/kepler/data\\_release.html](https://archive.stsci.edu/kepler/data_release.html).

tained and the details of the pipeline processing. The notes provide a summary of flight system events that affect the quality of the data and the performance of the pipeline for each data release. The Data Release Notes, along with other *Kepler* documentation, are located at MAST.

#### **1.2.4 *Kepler* Archive Manual (KDMC–10008)**

Data from the *Kepler* Mission are archived at MAST, which serves as NASA’s primary archive for ultraviolet and optical space-based data. The *Kepler* Input Catalog (KIC – Brown et al., 2011), processed light curves, and target pixel data are all accessed through MAST. The *Kepler* Archive Manual describes data products, file formats, and the functionality of *Kepler* data access. The archive manual can be accessed from the MAST *Kepler* page: <https://archive.stsci.edu/kepler/documents.html>.

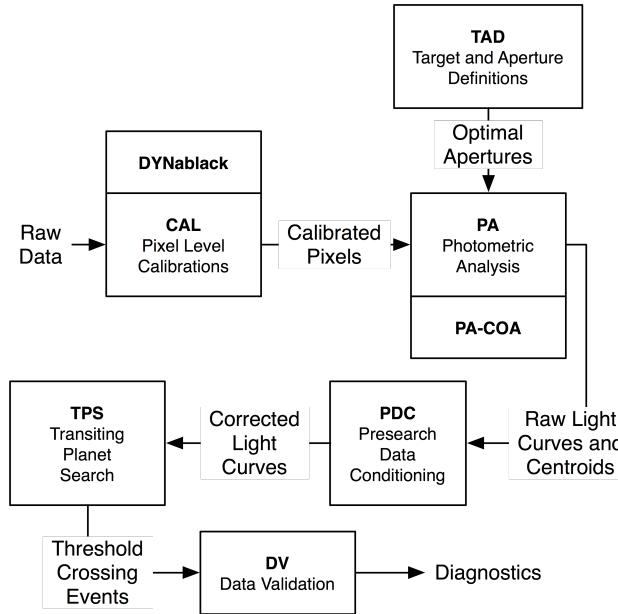
#### **1.2.5 *Kepler* Input Catalog**

While not a document per se, the KIC is an important reference and data source for those analyzing and interpreting data from the *Kepler* Mission. Prior to launch, the Stellar Classification Program (SCP) made multi-color observations of the entire *Kepler* FOV to determine the basic stellar properties for use in target selection (Batalha et al., 2010). The results of the SCP effort were collected into the KIC, which also federated information from a number of other existing catalogs. The KIC contains an assigned ID number, known as the KepID (sometimes KIC #), coordinates, photometry and stellar classification parameters. The Stellar Properties Working Group operated throughout the *Kepler* Mission to improve the properties of the target stars and other stars in the FOV and made several updates to the KIC (Huber et al., 2014). The MAST hosts KIC and an interactive query form at <https://archive.stsci.edu/kepler/kic.html>.

### **1.3 Document Organization – From Pixels to Planets**

The organization of this document first provides an overview of the SOC and its operations in Chapter 2, and then in Chapter 3 describes how the pixels are selected for storage and downlink onboard the *Kepler* spacecraft for target stars. The remainder of the document follows the progress of the raw data through the SOC Science Data Processing Pipeline as they are transformed from original pixel measurements to calibrated pixels, light curves and centroids, and ultimately searched for transiting planet signatures, also called “threshold crossing events” (TCEs). These TCEs are subjected to a number of diagnostic tests and are fitted with physical models to provide metrics and initial planetary parameters to the Threshold Crossing Event Review Team (TCERT), which then creates

Kepler Objects of Interest (KOI) and dispositions them using these diagnostics as well as other post-pipeline generated diagnostics.



**Figure 1.1** Data flow diagram of the *Kepler* Science Data Processing Pipeline. The journey of the raw data, from pixels to potential planets, is traced through the several components of the pipeline. See the text for a full explanation.

There are six major science modules in the *Kepler* SOC Science Data Processing Pipeline, as indicated in Figure 1.1. The Dynablock module (see Chapter 4) analyzes FFIs to identify rolling band artifacts and to provide updates to thermally-sensitive video crosstalk coefficients used in the pixel-level calibrations discussed in Chapter 5. The Photometric Analysis (PA) module, described in Chapter 6, identifies and removes the background flux and cosmic rays and, for each target star, sums the background-corrected pixels to formulate a brightness estimate and also measures the centroid offsets. In SOC 9.3, the optimal apertures provided by the Target and Aperture Definitions (TAD) module are replaced with optimal apertures that take the measured photometric precision and the actual image data into account. These new apertures are provided by the Create Optimal Apertures (PA-COA) component internal to PA, which is detailed in Chapter 7. The Pre-search Data Conditioning (PDC) module, detailed in Chapter 8, applies a multi-bandpass Bayesian maximum *a posteriori* (MAP) approach to identifying and removing instrumental signatures and systematic errors from the light curves while preserving intrinsic stellar photometric variations. These systematic error-corrected light curves are subjected to an adaptive, wavelet-based matched filter in the Transiting Planet Search (TPS) component (see Chapter 9) to search for transit-like features. Such features are passed to the Data Validation (DV) module, which fits limb-darkened transit models to the transit-like features (see Chapter 12) and constructs a suite of diagnostic tests

to make or break confidence in the transit-like features (see Chapter 11). The statistical bootstrap algorithm, one of the diagnostic tests performed in DV, is described in Chapter 10.

## Bibliography

- Akeson, R. L., Chen, X., Ciardi, D., et al., 2013. “The NASA Exoplanet Archive: Data and Tools for Exoplanet Research,” PASP, 125, 989
- Batalha, N. M., Borucki, W. J., Koch, D. G., et al., 2010. “Selection, Prioritization, and Characteristics of Kepler Target Stars,” ApJL, 713, L109
- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A., 2011. “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” AJ, 142, 112
- Huber, D., Silva Aguirre, V., Matthews, J. M., et al., 2014. “Revised Stellar Properties of Kepler Targets for the Quarter 1-16 Transit Detection Run,” ApJS, 211, 2
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)
- Van Cleve, J. E., Christiansen, J. L., Jenkins, J. M., et al. 2016. Kepler Data Characteristics Handbook (KSCI-19040-005) (Moffett Field, CA: NASA Ames Research Center)



## CHAPTER 2

---

# OVERVIEW OF THE SCIENCE OPERATIONS CENTER

JON M. JENKINS

NASA Ames Research Center, Moffett Field, CA 94035

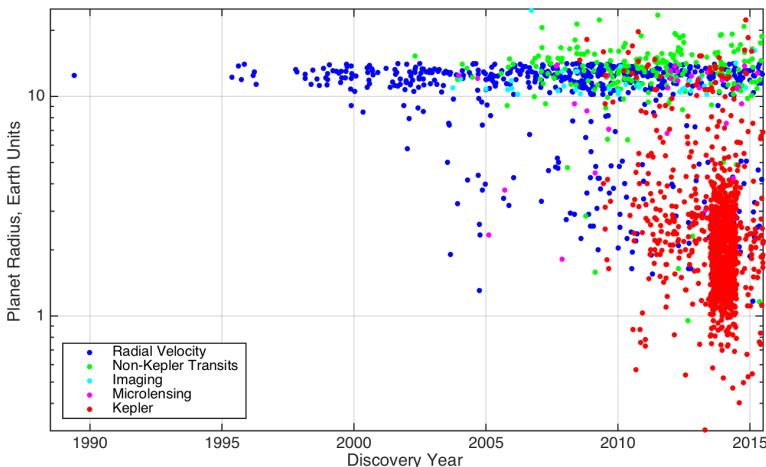
**Abstract.** The *Kepler* telescope launched into orbit in March 2009, initiating NASA’s first mission to discover Earth-size planets orbiting Sun-like stars. *Kepler* simultaneously collected data for  $\sim 165,000$  target stars at a time over its four-year mission, identifying over 4700 planet candidates, over 2300 confirmed or validated planets, and over 2100 eclipsing binaries. While *Kepler* was designed to discover exoplanets, the long-term, ultra-high photometric precision measurements it achieved made it a premier observational facility for stellar astrophysics, especially in the field of asteroseismology, and for variable stars, such as RR Lyrae. The *Kepler* Science Operations Center (SOC) was developed at NASA Ames Research Center to process the data acquired by *Kepler* from pixel-level calibrations all the way to identifying transiting planet signatures and subjecting them to a suite of diagnostic tests to establish or break confidence in their planetary nature. Detecting small, rocky planets transiting Sun-like stars presents a variety of daunting challenges, including achieving an unprecedented photometric precision of  $\sim 20$  ppm on 6.5-hour timescales, and supporting the science operations, management, processing, and repeated reprocessing of the accumulating data stream. This chapter describes how the design of the SOC meets these demanding requirements, discusses the architecture of the SOC and how the Pipeline is operated and run on the NASA Advanced Supercomputing (NAS) Division’s Pleiades supercomputer, and summarizes the most important pipeline features addressing the multiple computational, image, and signal processing challenges posed by *Kepler*.

**Keywords:** data processing; high performance computing; software architecture; astronomy: extrasolar planets; astronomy: astrophysics

### 2.1 Introduction

Until quite recently, the prospect of finding other worlds orbiting stars other than our own Sun was viewed as the province of science fiction. 51Peg b, the first exoplanet orbiting a normal, main-sequence hydrogen-burning star was discovered only in 1995, ushering in a new era of astronomy (Mayor & Queloz, 1995). This “hot Jupiter” was found in its 4.2-day period orbit at a mean separation of only 0.05 AU from its G2V host star by means of Doppler reflex velocity observations. In the early days of exoplanets, most were discovered from the ground

by such radial velocity observations, and most were comparable in mass to our solar system's Jupiter. Advances in technology and observational strategies have allowed the sensitivity to smaller-mass planets to improve at an astonishing rate of a factor of  $\sim 2.4$  each year as indicated in Figure 2.1.<sup>1</sup> As of October 2016,  $\sim 600$  exoplanets have been discovered via radial velocity.<sup>2</sup> Despite significant advances in the sensitivity of radial velocity searches, this technique strongly favors the detection of planets much more massive than Earth in most cases. Exceptions include small, cool M stars, with masses from 0.06 to 0.6 that of the Sun, for which the radial velocity signature of near Earth-mass planets is detectable. The most spectacular example is Proxima b (Anglada-Escudé et al., 2016), a 1.1 Earth-mass planet in an 11-day period orbit of Proxima, a small, cool M star only  $\sim 12\%$  the size of the Sun and only  $\sim 1.3$  parsec from the Earth.



**Figure 2.1** A history of exoplanet discoveries plotted in terms of the estimated planetary radius vs. discovery year. The discovery method is indicated by the color of the points, with blue indicating planets discovered via radial velocity, green indicating non-*Kepler* transiting planets, cyan indicating planets discovered via direct imaging, magenta indicating planets discovered via gravitational microlensing, and red points indicating *Kepler* discoveries. Note that the discovery year for each planet has been randomly dithered to reduce over-plotting.

By contrast, transit photometry searches did not get underway in earnest until after the radial velocity discovery of the first transiting planet, HD209458b, was made in 2000 (Henry et al., 2000; Charbonneau et al., 2000). In transit photometry, exoplanets are discovered by measuring the minute diminution of light that occurs when a planet crosses the face of its star. Transit photometry has the advantage of being sensitive to the areas of planetary discs relative to their stars, where the Earth's projected area is only 121 times smaller than that of Jupiter (compared to the corresponding mass ratio of 318). Nevertheless, the technical

<sup>1</sup>For planets with masses greater than Earth,  $M \approx R^{-1.3}$  (Wolfgang et al., 2016).

<sup>2</sup>Exoplanet discovery statistics are taken from the online table at <http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=planets>.

challenges of transit photometry, coupled with the difficulties of making sufficiently sensitive long-term, continuous observations from the ground delayed the achievement of the promise of this technique to find small rocky planets until the *Kepler* Mission was on orbit.

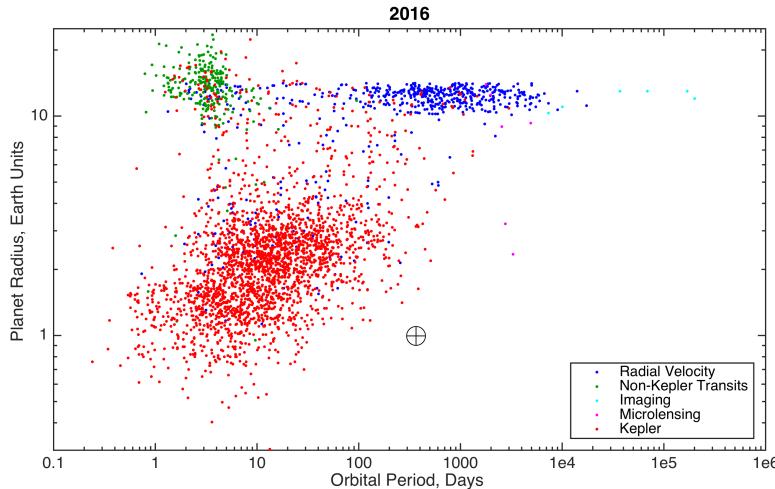
The *Kepler* Mission was designed to discover Earth-size planets orbiting Sun-like stars through transit photometry: observing the small diminution of light that occurs when a planet crosses the face of its star from the observatory's point of view (Borucki et al., 2010). The amplitude of the planetary signal is minute,  $\sim 100$  ppm, and lasts from  $\sim 1$  hour to about half a day. This signature must be recognized against a variety of noise sources that are often much larger in amplitude, including instrumental effects, such as shot noise and thermally-induced focus and pointing variations, and intrinsic stellar variability, including star spots and granulation noise. The transits repeat once per orbital period with an unknown phase, necessitating observations that can be carried out as long and as continuously as possible. In addition, most orbital configurations inhibit the observation of transits, as only a small fraction of possible inclination angles allows the planet to cross the face of the star from our point of view.

The *Kepler* Mission rose to these challenges with a 0.95-m aperture telescope that launched into orbit in March 2009 to conduct nearly continuous observations of up to 170,000 stars at a time in a single 116-square degree field of view (FOV) over a four-year mission. *Kepler* acquired data at 29.4-minute intervals for all target stars called long cadence (LC) targets and at 1-min intervals for up to 512 target stars at any given time called short cadence (SC) targets, and the resulting flux time series were typically  $>90\%$  complete. The observations were organized into seventeen 93-day “quarters” to rotate the telescope by  $90^\circ$  every three months in order to keep the sunshade and the solar arrays properly oriented (Haas et al., 2010).<sup>3</sup> As of October 2016, 2679 planets have been discovered via transits, including 2330 planets discovered by *Kepler*, as illustrated in Figure 2.2.

The characteristics that made *Kepler* such a successful planet hunter also made it a near perfect stellar variability observation machine, providing important science results across a diverse set of stellar phenomena, including asteroseismology, gyrochronology, spot modulation, super flares, eclipsing binaries, “heart-beat stars,” and relativistic boosting.

Undoubtedly, the success of *Kepler* was enabled by the exquisite instrument with its 95-megapixel camera and the benign orbit it occupies. Of equal importance is the Science Data Processing Pipeline, which needed to keep up with the accumulating data volume, extract photometry at the 20-ppm level with a raw precision of  $\sim 2\%$ , and permit timely reprocessing of the dataset as the pipeline evolved and its sensitivity improved.

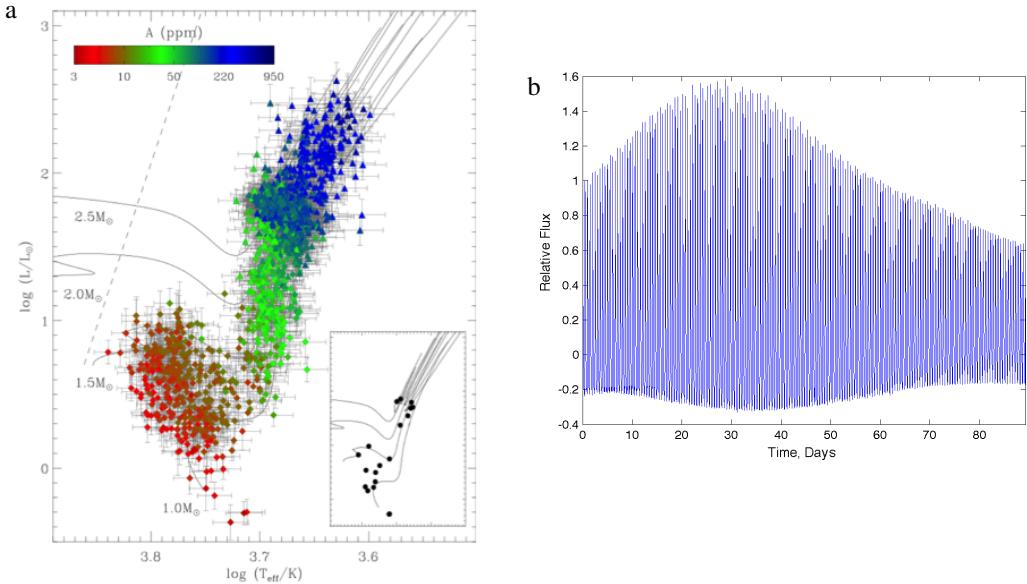
<sup>3</sup>The first quarter, Q1, was only 34 days long due to the launch date and commissioning period. The last quarter, Q17, was only 31 days long, due to the mission-ending loss of reaction wheel #4, and contained a 10-day rest period to attempt to increase the lifetime of this reaction wheel.



**Figure 2.2** Radius vs. orbital period for all known exoplanets as of October 2016. The discovery method is indicated by the color of the points, with blue indicating planets discovered via radial velocity, green indicating non-*Kepler* transiting planets, cyan indicating planets discovered via direct imaging, magenta indicating planets discovered via gravitational microlensing, and red points indicating *Kepler* discoveries. Clearly, the most successful exoplanet survey to date in terms of discovering small planets ( $R_p < 2R_{\oplus}$ ) in the habitable zone of their star is the *Kepler* Mission.

The 1-min short cadence interval permitted observations of pressure- or p-mode oscillations of solar-like stars, which have typical oscillation periods much less than a half hour. Despite being designed exclusively for the purpose of detecting minute drops in brightness corresponding to transit events, *Kepler* has proven itself adept at revealing stellar variability over a huge dynamic range from p-mode oscillations of  $\sim$ 10 ppm and transit signatures of  $\sim$ 100 ppm, to oscillations of RR Lyrae stars, which can nearly double their brightness every half day. Figure 2.3 illustrates the large dynamic range ( $10^{5.8}$ ) of photometric features identified in *Kepler* light curves with SOC 9.3.

The *Kepler* SOC was developed at NASA Ames Research Center over a 12-year period of time beginning in 2004 and continuing through the primary mission and well into the extended mission. Three principle factors stimulated significant research and development of new algorithmic approaches for virtually every module of the science data processing pipeline: 1) the stellar variability of the main-sequence stars in *Kepler*'s FOV proved to be twice as large as expected, based on long-term observations of the Sun (Gilliland et al., 2011, 2015), 2) instrumental effects caused both by radiation damage and by electronic image artifacts triggered an overabundance of false alarms and threatened to overwhelm the system (Caldwell et al., 2010; Coughlin et al., 2014; Mullally et al., 2015), and 3) the interplay of the intrinsic stellar signatures and instrumental signatures required the development of more sophisticated approaches to identifying and removing systematic errors than were available in the original pre-launch pipeline design (Jenkins et al., 2012).



**Figure 2.3** a: A Hertzsprung–Russell diagram displaying  $\log$  luminosity vs.  $\log$  effective temperature for 15,000 stars exhibiting p-mode oscillations observed by *Kepler* (Huber, 2016). The points are colored by the amplitudes of the stellar oscillations, which vary from 3 ppm to  $\sim 3600$  ppm, illustrating that the amplitudes vary with the mass and size of the star. The inset shows similar results for  $\sim 20$  stars obtained prior to 2008. b: Light curve for the RR Lyr star KIC 7671081 exhibiting amplitude modulation via the Blazhko effect. The SOC 9.3 pipeline greatly reduces the distortions of the intrinsic astrophysical signals in the light curves of high amplitude variable stars compared to SOC 7.0 and earlier codebases.

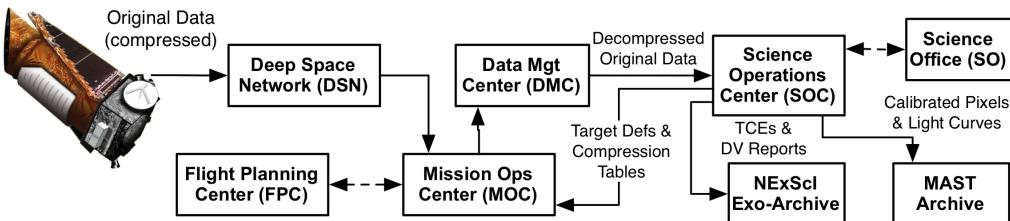
As the pipeline evolved, the data needed to be reprocessed, and this, too, was a challenge. While the 700-node computer cluster used to process the *Kepler* data was able to keep up with the data as it was downlinked every month, it could not reprocess the accumulating data record in a reasonable amount of time. This motivated the SOC to develop new software infrastructure in order to be able to routinely process and reprocess data on the NAS Division’s Pleiades supercomputer.<sup>4</sup>

## 2.2 The *Kepler* Mission Ground Segment

The *Kepler* ground system is illustrated in Figure 2.4, and includes the elements necessary to manage the spacecraft and instrument, collect, analyze, and interpret the science data, and deliver it to a permanent archive.

The *Kepler* spacecraft transmits the compressed, raw science and engineering data through the Deep Space Network (DSN) on monthly intervals to the Mission Operations Center (MOC) in Boulder, CO, which collects the necessary

<sup>4</sup>Pleiades currently has 227,808 computer cores and 828 TiB of memory (<http://www.nas.nasa.gov/hecc/resources/pleiades.html>).



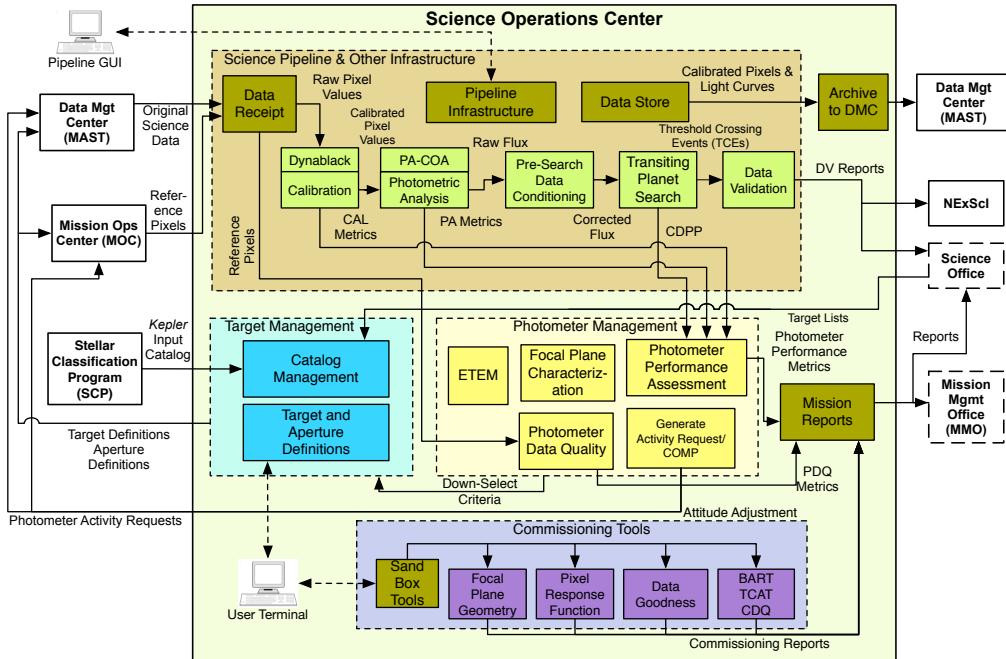
**Figure 2.4** Organization of the *Kepler* ground segment, illustrating the flow of science data from the spacecraft and across the system. This system manages the *Kepler* science instrument to collect the science data and the information necessary to carry out the *Kepler* Mission.

ancillary engineering data (AED)<sup>5</sup> and forwards them along with the science data to the Data Management Center (DMC) at the Space Telescope Science Institute (STScI) in Baltimore, MD. The DMC decompresses the science data and packages them and the ancillary engineering data into FITS files that are sent to the SOC. The SOC processes the data and then exports calibrated pixels and flux time series for every target star to the Mikulski Archive for Space Telescopes (MAST) at STScI. The SOC also exports target lists, stellar properties used by the pipeline, and focal plane models used in calibrating the data to the MAST. The SOC sends target definitions that specify the pixels to be stored and downlinked by the spacecraft for every target star, compression tables, and data collection parameters to the MOC for uplink to the spacecraft. The results of the planet search are sent to NExSci's exoplanet archive (Akeson et al., 2013) in Pasadena, CA. The Science Office at NASA Ames Research Center supports the data processing at the SOC and provides calibration files, target lists, and stellar parameter updates to the SOC. The FPC at Ball Aerospace & Technologies Corporation in Boulder, CO manages the science instrument.

## 2.3 The Science Operations Center

The SOC processing system consists of several elements: 1) a pipeline infrastructure coded in Java that ingests the science data, controls the science processing, and writes the archive data products to files in the archive file format, 2) the science pipeline itself, 3) a target management system that contains a catalog of target and field stars and their characteristics and that identifies the pixels of interest for each target star and associated on-chip collateral data, and 4) a suite of commissioning tools used to obtain or validate various calibration models and pre-flight instrument characterizations. Figure 2.5 shows the high-level architecture of the SOC software system.

<sup>5</sup> AED is a subset of engineering data that bears on the state and performance of the instrument, such as the temperature of the focal plane and reaction wheel speeds.



**Figure 2.5** Architecture diagram of the *Kepler* Science Operations Center indicating the 23 major components comprising the science pipeline, target management, commissioning tools, and photometer management functions.

### 2.3.1 Software Infrastructure

**2.3.1.1 Pipeline Infrastructure (PI)** The pipeline infrastructure provides automated distributed processing of science data and sequencing of modules based on the results of previous modules (Klaus et al., 2010a,b). Features include a customizable unit-of-work that controls how the data are distributed across the cluster, a configuration management and versioning system for algorithm parameters and pipeline configurations, and a graphical user interface for the configuration, execution, and monitoring of pipeline jobs. PI provides scalability for running the pipeline on a developer workstation, a large cluster of computing nodes, as well as on the NASA Pleiades supercomputer.

PI is designed to be a generic, reusable platform suitable for developing science processing and analysis pipelines. PI provides a plug-in architecture for pipeline modules, parameter definitions, and unit-of-work definitions, and contains no *Kepler*-specific code. PI also provides a generic mechanism for running any pipeline module on large computing clusters that support the PBS (Portable Batch System) interface. This allows nearly all *Kepler* science data to be processed routinely on NASA's Pleiades system.

**2.3.1.2 Data Receipt (DR)** This module provides data ingestion and automated pipeline launch capabilities and is divided into two main components: 1) a

generic layer that watches for new files, dispatches the proper handler, and launches pipelines and 2) a plug-in layer for specific data types.

**2.3.1.3 *Mission Reports (MR)*** MR provides a web-based interface to a library of reports concerning the pipeline and its processes that can be generated on the fly. These reports are used extensively by the operations personnel to help manage the science processing.

**2.3.1.4 *Data Store (DS)*** The *Kepler* DS is a transactional database management system for arrays, sparse arrays, and binary data (McCauliff et al., 2010). DS consists of a custom array data base (ADB) and an Oracle database. The data volume of the 4-year *Kepler* dataset is  $\sim$ 32 TiB. The working copies of the *Kepler* data sets, including intermediate files used for quality analysis and testing occupy  $\sim$ 500 TiB, and reside on an NFS server.

**2.3.1.5 *Archive (AR)*** AR generates the files in the format archived to MAST and made available to the science team and greater astronomical community. The *Kepler* archival products include calibrated pixels, simple aperture photometry, systematic-error-corrected photometry, astrometry (centroids), and associated uncertainty estimates. Target pixel files contain the pixel data, both original and calibrated, for each target organized as image data. These files also include information about sky background flux and cosmic ray hits detected by the pipeline. The transit-like features identified by the transit search and the results of DV's diagnostic tests are archived as XML files along with PDF reports to the NASA Exoplanet Archive managed by NASA's Exoplanet Science Institute. These PDF files contain a wealth of information regarding each transit-like signature.

**2.3.1.6 *Support Libraries (SLIB)*** The PI depends on a set of support libraries that have been customized and augmented specifically for the SOC. Most of this software is written in Java, but there is some MATLAB software in this collection to provide, for example, the capability to retrieve data and model information from the Data Store from an interactive MATLAB session on a SOC development workstation. Much of this software underlies the automated test suites that are exercised on a nightly basis to check the codebase integrity.

**2.3.1.7 *Target Management*** The *Kepler* target management functionality consists of two components:

1) Catalog Management (CM) contains the *Kepler* Input Catalog (KIC) provided by the Stellar Classification Program (Brown et al., 2011) and subsequent updates to the catalog (Huber et al., 2014). CM contains the characteristics of the target stars and background field stars, such as location (right ascension, declination), effective temperature, surface gravity, radius, mass, and proper motion, which are necessary to select target stars and to identify the pixels to be used in measuring the brightness of each target star; and

2) Target and Aperture Definitions (TAD), which formulates the target definitions specifying which pixels need to be stored and downlinked by the *Kepler*

spacecraft. TAD also formulates the 1024 mask definitions used to capture the pixels of interest of the ensemble of target stars. The associated sub-module, Compute Optimal Apertures (COA), predicts the pixels of interest for extracting photometric measurements from the CCD images for each target star (Bryson et al., 2010c). *Kepler* had very tight margins for the pixel data stored onboard and returned to the ground: only  $\sim$ 6% of the pixel data could be stored onboard for later downlink. TAD is described in detail in Chapter 3.

### 2.3.2 Photometer Management

This suite of software contains the calibration models used by the science pipeline as well as modules that monitor the performance of the photometer.

**2.3.2.1 Photometer Performance Assessment (PPA)** This component assesses the health and performance of the instrument based on the science data sets collected each month, identifying out-of-bounds conditions and generating alerts (Li et al., 2010). PPA is implemented with multiple pipeline modules: PPA Metrics Determination (PMD), PMD Aggregator (PAG), and PPA Attitude Determination (PAD). Various metrics are generated as the science pipeline processes the science data, including photometric precision, brightness, black level, background flux, smear level, dark current, cosmic ray counts, outlier counts, centroids, reconstructed attitude, and the difference between the reconstructed and nominal attitudes. These metrics are tracked and trended by PPA and the results are persisted to the Data Store and are used to generate a PDF report that is available for review by project personnel. PAD combines the astrometric data collected for each CCD readout channel to construct a high-fidelity record of the pointing history for each CCD channel for each 29.4-min data collection interval, comparing them against the nominal attitudes. The output of PPA is used to identify and set data anomaly flags required for processing the science data to produce the archive data.

**2.3.2.2 Photometer Data Quality (PDQ)** PDQ provides a quick look assessment of the health and performance of the instrument through data downlinked at X-band twice-weekly called “reference pixels” (Chandrasekaran et al., 2010). A small amount of *Kepler* science data consisting of no more than 96,000 Long Cadence (LC) pixels ( $\sim$ 200 target stars) are collected once per day and stored in non-volatile memory for downlink at the twice-weekly X-band contacts with the DSN. This provides frequent updates on the status and health of the photometer and assesses the validity of the spacecraft’s attitude for science operations following a return to science attitude. If any of the target stars are more than 0.1 pixels ( $\sim$ 0.4'') from their desired location in the first reference pixel file, the SO authorizes a pointing tweak to correct the pointing. PDQ tracks and trends various aspects of the photometer and provided the first indication of the loss of module 3 failure in January 2010.

**2.3.2.3 *End-To-End Model (ETEM)*** ETEM is a suite of software that generates synthetic flight-like data for *Kepler* with a high degree of fidelity, including matching the formats of the science data at each ground segment interface, from the solid state recorder (SSR) onboard the spacecraft, through the MOC and the DMC (Jenkins et al., 2004; Bryson et al., 2010a). ETEM was indispensable in testing the entire *Kepler* ground segment as well as for designing, implementing, and testing the SOC. ETEM simulates the astrophysics of planetary transits, stellar variability, background and foreground eclipsing binaries, cosmic rays, and other phenomena.

**2.3.2.4 *Focal Plane Characterization (FC)*** This module consists of a set of database tables, persistence classes, and associated handling code that manages the calibration models used to process data and manage target definitions (Allen et al., 2010). These include the 2-D black (bias voltage) image, the pixel-level gain model, and the pixel response function (PRF). One of the most critical model sets maintained in FC is the model describing the mapping from celestial coordinates to focal plane coordinates (pixels) called RaDec2Pix.

**2.3.2.5 *Compression (COMP)*** Pixel data compression tables are generated by two components named the Huffman Generator (HGN) and the Huffman Aggregator (HAG) modules. The data compression scheme involves three steps: 1) re-quantizing the data so that the quantization noise is approximately a fixed fraction of the intrinsic measurement uncertainty (which is dominated by shot noise for bright pixels), 2) taking the difference between each re-quantized pixel value and a baseline value that was updated once per day, and 3) entropic encoding via a length-limited Huffman table (Jenkins & Dunnuck, 2011). The compression scheme compressed the original 23-bit words to typical lengths of 4.5-5 bits per pixel measurement on average for a compression ratio of  $\sim$ 5:1. This level of compression allowed for  $>$ 66 days of data to be stored on the SSR, thereby decreasing the time required for DSN contacts and increasing robustness against missed DSN passes.

### 2.3.3 Science Pipeline

The Science Pipeline calibrates the original data from *Kepler* and produces the archival data products. It conducts the transit search and constructs diagnostics used to prioritize and rank the planetary candidates for follow-up observations.

**2.3.3.1 *Calibration (CAL)*** This module operates on original spacecraft pixel data to remove instrumental artifacts such as smear from the shutterless readout. Traditional CCD data reduction is performed (removal of instrument/detector effects such as bias, dark current and flat field), in addition to *Kepler*-specific pixel-level calibrations, such as the Local Detector Electronics (LDE) undershoot correction (Quintana et al., 2010). The associated Dynablock (DYN) module analyzes Full Frame Images (FFIs – normally acquired once per month) to identify rolling

band artifacts that can trigger false positives and to update thermally sensitive Fine Guidance Sensor (FGS) crosstalk coefficients used in pixel-level calibrations by CAL. CAL operates on both SC and LC data to produce calibrated pixel flux time series, associated uncertainties, and metrics that are used in subsequent pipeline modules. See Chapter 4 and Chapter 5 for details of the DYN and CAL modules.

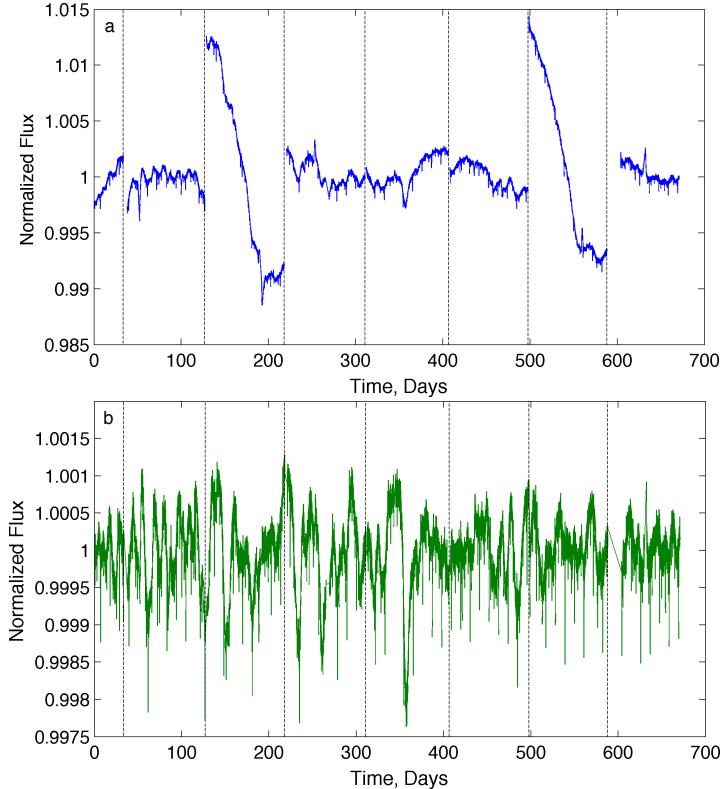
**2.3.3.2 Target and Aperture Definitions (TAD)** In the context of the science pipeline, TAD updates the photometric apertures based on the reconstructed pointing history obtained by centroiding a fiducial set of bright, unsaturated targets on each 29.4-min cadence.

**2.3.3.3 Photometric Analysis (PA)** PA measures the brightness of each target star on each frame. It also fits and removes background flux due to zodiacal light and the diffuse stellar background, identifies and removes cosmic rays from all target star apertures, and measures the photocenter or centroid of each target star on each frame. PA uses PRF-fitting to measure precisely the location of  $\sim 200$  bright, unsaturated target stars on each CCD channel in order to establish the pointing and focus of the photometer. This information is used along with image data to update the photometric apertures (Smith et al., 2016) using a new SOC 9.3 sub-component called Create Optimal Apertures (PA-COA), detailed in Chapter 7. The architecture and algorithms of PA are presented in Chapter 6.

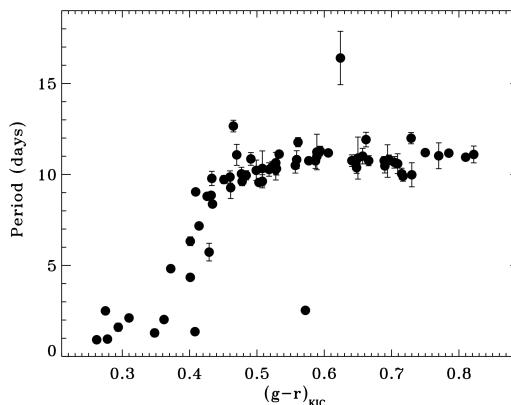
**2.3.3.4 Presearch Data Conditioning (PDC)** PDC performs a critical set of corrections to the light curves produced by PA, including the identification and removal of instrumental signatures caused by changes in focus or pointing and of step discontinuities that result occasionally from radiation events in the CCD detectors. PDC also identifies and removes isolated outliers and corrects the flux time series for crowding effects and for the fact that not all the light from a star can be captured by a finite aperture (Stumpe et al., 2014; Smith et al., 2012). PDC employs a multi-scale maximum *a posteriori* (MAP) approach to identify and remove systematic errors, allowing it to retain important astrophysical signals in the face of much larger instrumental effects, as illustrated by Figure 2.6. The architecture and algorithms of PDC are presented in detail in Chapter 8.

Figure 2.7 illustrates the results of analyzing light curves produced by PDC-MAP to measure the rotation periods of a set of stars in the open cluster NGC 6811, permitting the calibration of the theory of gyrochronology to  $\sim 1$  Gyr-old stars (Meibom et al., 2011).

**2.3.3.5 Transiting Planet Search (TPS)** TPS implements a wavelet-based, adaptive matched filter algorithm to detect signatures of transiting planets (Jenkins, 2002; Jenkins et al., 2010). TPS stitches the quarterly light curves together prior to searching for planets. TPS also provides estimates of combined differential photometric precision (CDPP), a key performance diagnostic for transit survey missions on the timescales of transits (Christiansen et al., 2012). This metric, used



**Figure 2.6** a. The original simple aperture photometry light curve for Kepler-20, a star hosting five transiting planets, including the first Earth-size planet discovered by *Kepler* (Gautier et al., 2012). b. The systematic error-corrected light curve produced by PDC using a MAP-based approach, showing good preservation of the transit signatures and rotational modulation of the star’s brightness by star spots. The vertical scales of these two panels are notably different: the vertical scale of the bottom panel is  $\sim 10\times$  smaller than that of the top panel.



**Figure 2.7** Stellar rotation periods measured for 71 dwarf stars in the  $\sim 1$ -Gyr-old open cluster NGC 6811 as a function of color, illustrating the strong relationship between color and rotation period for stars of the same age. PDC-MAP reliably preserves intrinsic variations on timescales  $< 20$  days. From Figure 4a in Meibom et al. (2011).

for both the *Kepler* Mission and the upcoming TESS Mission (Ricker et al., 2015), is necessary for estimating the completeness of a transit survey and for extrapolating the results to infer the intrinsic frequency of planets in the star sample. TPS is described in detail in Chapter 9.

**2.3.3.6 Data Validation (DV)** This component performs a suite of diagnostic tests on each transiting planet signature identified by TPS to make or break confidence in its planetary nature, as detailed in Chapter 11. These include a comparison of the depth of the even transits to the odd transits, an examination of the correlation of changes in the photocenter (centroid) of the target star to the photometric transit signature, a statistical bootstrap to assess confidence in the detection (see Chapter 10), difference image centroiding to rule out background sources of confusion, and a ghost diagnostic test to rule out optical ghosts of bright eclipsing binaries as the source of the transit-like features. These tests can determine if the transit signature is likely to be due to a background eclipsing binary whose diluted eclipses are masquerading as transits of a planetary body. DV also calls TPS to search the residual light curve for evidence of additional transiting bodies after fitting and removing the first planetary transit signature (see Chapter 12). This process is repeated until TPS fails to identify another transit signature or hits a programmable upper limit (9 in the run of SOC 9.3 for DR25).

### 2.3.4 Commissioning Tools

Several tools were developed and deployed specifically for the commissioning phase of the *Kepler* Mission.

**2.3.4.1 Focal Plane Geometry (FPG) and Pixel Response Function (PRF)** FPG and PRF were used to determine the detailed sky to pixel mapping coefficients and the shape of the point spread function (PSF), respectively, across each of the 84 individual CCD readout channels. The FPG coefficients include terms for pin-cushion distortion (Tenenbaum & Jenkins, 2010). PRF constructed five individual PRFs for each CCD readout area in order to capture non-uniformity in the focus and PSF (Bryson et al., 2010b). PRF models at intermediate locations are obtained by interpolation. The pipeline uses these model waveforms to select target apertures and to monitor the locations of the brightest, unsaturated 200 stars on each channel in order to reconstruct pointing and capture distortion due to focus changes.

**2.3.4.2 Pixel Overlay On FFIs (POOF)** This tool allows the user to retrieve *Kepler* FFIs and overlay the aperture masks from target tables on the images as well as information about the stellar targets themselves. This enabled the validation of target definition tables early in the *Kepler* Mission via manual inspection of the degree to which the selected target masks covered the target star images, and also identified deficiencies in the KIC, such as outdated celestial coordinates for the Lépine stars (Lépine et al., 2013). POOF is a tool that is still used to examine the FFIs to diagnose issues with time series pixel data.

**2.3.4.3 Data Goodness (DG)** The DG tool allows the user to examine diagnostic statistics on *Kepler* FFIs that were collected during commissioning but were not fully analyzed at the time. This tool is still in use to verify the quality of *Kepler* FFIs collected during normal science operations.

**2.3.4.4 BART, TCAT, and CDQ** These tools were designed specifically to monitor the behavior of certain electronic artifacts discovered pre-launch. The 2-D Black and Artifact Removal Tool (BART) detects and models temperature-dependent image artifacts in pixel data. BART provides insight into whether the photometer data are consistent with pre-launch expectations regarding temperature variations. The Temperature Coefficient Analysis Tool (TCAT) investigates the thermal variations of pixels affected by crosstalk caused by the Fine Guidance Sensors (FGS) used to control the spacecraft pointing. FGS Crosstalk is a significant source of *Kepler* instrument noise. Check Data Quality (CDQ) checks and analyzes the rms of data-fitting residuals and thermal coefficients produced by BART for the pixels in the collateral regions of the photometer.

**2.3.4.5 Sandbox Tools (SBT)** The sandbox tools allow *Kepler* personnel to make queries against the ADB on their own workstations.

### 2.3.5 Hardware

This section describes the hardware used to support the pipeline.

**2.3.5.1 Cluster Worker Machine** Cluster worker machines are used to serialize data inputs for MATLAB executables that run on Pleiades and to parse the outputs generated from those pipeline modules. Worker machines can run these MATLAB executables locally, but at a much smaller scale. This is done for less numerically intensive modules such as PPA and AR where the additional complexity of running on Pleiades is not warranted.

Cluster worker machines can be moved between different clusters in order to provide for some redundancy. We procured worker machines with 24 cores (48 including hyper threads) and 768 GiB of RAM. Temporary task file storage is on a local NFS. This means local storage on the worker machine is not a bottleneck for either storage capacity, performance, or availability. Two 10 GiB Ethernet network interfaces are present on the worker machines.

### 2.3.6 Cluster Datastore Machine

Each cluster also has a dedicated datastore machine that has a similar specification to the cluster worker machines with the addition of two 8 GiB fiber channel host bus adapters. Oracle and ADB share this machine. The *Kepler* Storage Area Network (SAN) has a storage capacity of  $\sim$ 500 TiB and a theoretical transfer rate limit of  $2 \text{ GiB sec}^{-1}$ , which is sufficient to copy the entire contents of the storage array in about 10 minutes. Practically, other parts of the architecture are

the limiting factor as the number of concurrent I/O operations is limited by the number of disks rather than by the network.

ADB is allocated 64 GiB of RAM with Oracle using the remainder. The version of ADB used by *Kepler* uses most of its memory to cache b-tree indices and for temporary buffers. Index blocks are used to locate array blocks on disk; this scales with the number of independent arrays in working memory, which is typically largest during processing for CAL. The total memory required for *Kepler* is  $\sim$ 32 GiB but 64 GiB is available. Oracle tends to be bottlenecked on disk I/O rather than processor power.

**2.3.6.1 Data Storage** Data storage is handled via a SAN, a dedicated network for the transmission of blocks of data to and from datastore machines. We use a storage array with  $\sim$ 200 7.2k RPM hard disks. The storage array presents the view of one or more virtual block devices to each host known as volumes or logical unit numbers (LUN). Each volume is in fact a combination of disks placed in a RAID 6+0 configuration. This allows for each LUN to be striped across all the drives in the array and so each can access the full number of I/O operations. The storage array can provide approximately 15k I/O operations per second. A volume can also be snapshot, which is a point-in-time copy of a base volume. Modifications to snapshots have a copy-on-write feature, which means space is only allocated for modifications. Failed disks can be replaced with reserve space on the remaining working disks. At a minimum, two disks can fail without a loss of data. In practice, many more disks have failed without data loss.

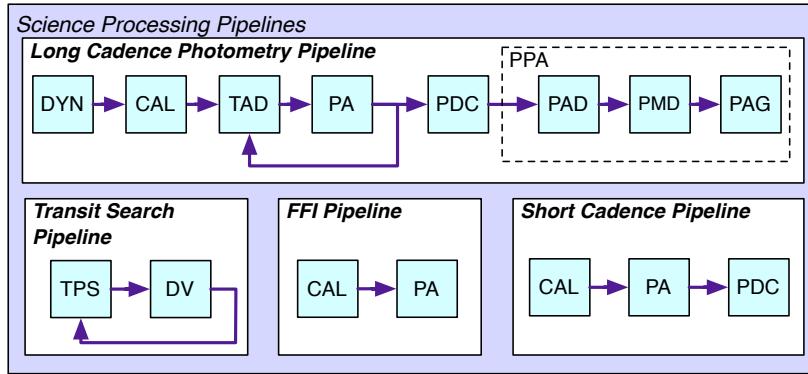
## 2.4 Running the Pipeline

The SOC Science Processing Pipeline is actually configured as multiple pipeline segments based on the dataset types that they process and the frequency at which they run, as indicated in the typical pipelines depicted in Figure 2.8. Each pipeline is a directed graph of pipeline modules.

Organizing these processing steps as separate pipelines provides flexibility without complicating the pipeline configuration. This flexibility also allows other scenarios to be implemented. The set of pipeline segments and the modules they contain are completely configurable using the pipeline GUI. The set of available modules (the module library) is also configurable. This architecture enables modules to be easily updated, added, or removed without code changes (other than to the modules themselves).

### 2.4.1 Unit of Work

Because the pipeline algorithms can be very computationally intensive and due to the large data volumes involved, the pipeline can be run on multiple machines. To this end, the pipeline can run on a cluster of local worker machines or a set of



**Figure 2.8** There are four major *Kepler* pipelines. Top panel: The photometry pipeline for LC data indicating that TAD is run twice, once before the pointing is reconstructed by PA and once afterwards. Only LC data are processed by PPA for attitude determination (PAD), for metric determination (PMD) and for aggregation of the metrics (PAG) and generation of the reports. Left bottom panel: The transit search pipeline for the LC data, indicating that DV can call TPS multiple times. Middle bottom panel: The FFI pipeline. Right bottom panel: The SC data pipeline.

remote machines on the NAS Pleiades supercomputer. When a new pipeline is launched, the work is divided into units that can then be distributed to individual worker machines. The unit-of-work (UOW) can be configured to bin the input data by cadence, CCD output, and/or target. The design goal is to use the UOW as a tuning knob to maximize concurrency. This knob is adjusted based on how many machines are available and how long a UOW takes to process.

For execution on Pleiades, a UOW is further broken into subtasks so that work can be distributed across the tens of thousands of nodes and cores in Pleiades. Additional tuning parameters control how many subtasks can execute on each node so as to best take advantage of the memory and cores present in each node. Table 2.1 lists units of work and subtasks for several pipeline modules.

**Table 2.1** Units of work to the subtask level for various pipeline modules.

Pipeline	
Module	Binned by
CAL	cadence interval, CCD output, CCD row(s)
PA	month (SC) or quarter (LC), CCD output, individual targets
PDC	month (SC) or quarter (LC), CCD output
TPS	individual targets
DV	individual targets

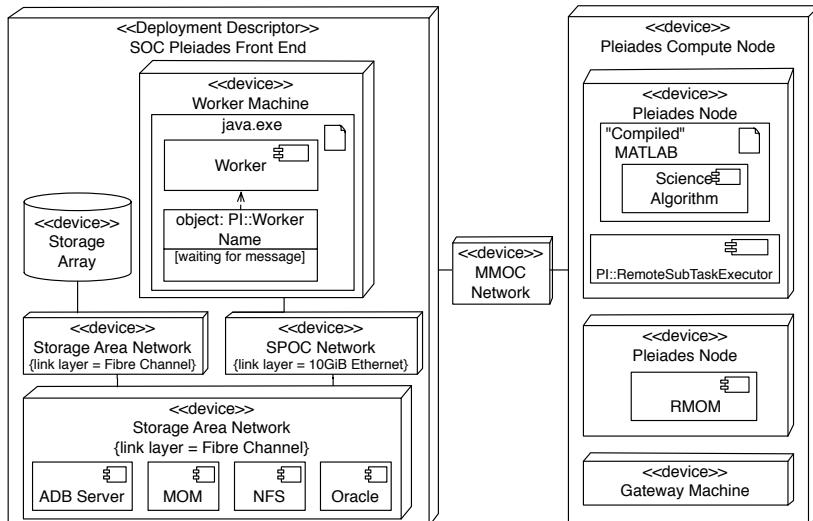
## 2.4.2 Coordination of Work

**2.4.2.1 Local Cluster** When a new pipeline is launched, a pipeline task is created for each UOW for each module in the pipeline. Pipeline tasks are scheduled for asynchronous execution using a distributed message queue also known as Message Oriented Middleware (MOM). At the start of execution, a message is placed

on the queue for each pipeline task. Once the messages are on the MOM queue, the next available worker machine will pull the next message off the MOM queue and execute the pipeline task corresponding to that message. Any worker is able to execute any pipeline task because each worker machine has access to all of the science modules and the pipeline infrastructure services. This design allows worker machines to be easily added for increased processing throughput.

#### 2.4.3 Remote Execution

Remote execution is distinguished from local execution by the use of third-party authentication and connection tools in order to execute pipeline tasks on Pleiades. In this case the pipeline worker processes remain local and files are transmitted over secure shell (ssh) via the Multi Mission Operations Center (MMOC) network. A remote queuing system (RMOM) allocates super computer nodes to subtasks. Pipeline modules can generate a dependency graph that expresses the dependencies between subtasks, such as the fact that the image motion information needs to be generated by PA prior to assigning the final photometric apertures via TAD. The RMOM obeys this dependency graph and so as many independent subtasks are run on at least as many available processing nodes (Figure 2.9). There are additional parameters that determine the number of concurrent subtasks that can execute on a processing node. This is usually limited by the memory-to-core ratio of the type of subtask being executed.



**Figure 2.9** Pipeline deployed on Pleiades. In this deployment, the local cluster is used to generate inputs and outputs for subtasks on Pleiades. Remote processes manage the execution of science algorithm implementations. This can scale up to tens of thousands of independent subtasks.

#### 2.4.4 Triggers

New pipeline instances are launched using pipeline triggers that associate pipeline parameters with specific pipeline instances. These triggers are part of the pipeline configuration and are created by the pipeline operator using the pipeline GUI. Triggers can also be used to launch pipelines manually, as in the case of reprocessing, or automatically on a particular schedule or when the input data become available. These data-available triggers allow the various pipeline types to be chained together so that complete, end-to-end science processing can be automated. For example, the photometry pipeline can be configured to run when new data are delivered from the spacecraft.

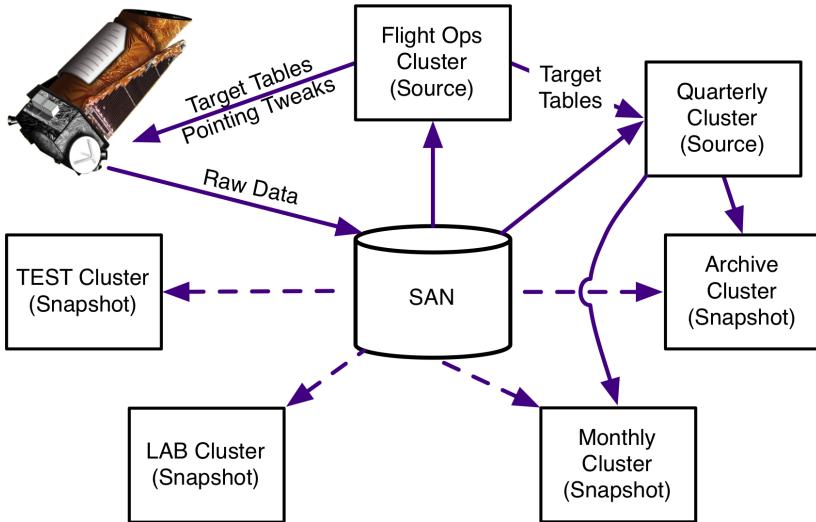
#### 2.4.5 Data Accountability

Data accountability is an integral and crosscutting feature of the SOC design and is designed into the pipeline infrastructure, datastore, data receipt, and each science pipeline module. Tracked data are assigned a unique originator ID that determines its origin. PI manages the sets of parameters used for each pipeline module. These parameter sets are locked into an immutable state once a pipeline trigger has been fired. Pipeline instance IDs ensure the actual parameter values used to process any piece of data are recorded and recoverable.

#### 2.4.6 Operating the Pipeline

The data processing tasks are distributed over four operations clusters: Flight Ops, Quarterly Ops, Monthly Ops, and Archive, and two test clusters: TEST and LAB. Note that all data are stored in the SAN, which is partitioned to support all six clusters. All original spacecraft science and engineering data, models, algorithm parameters and configuration settings used for archival tasks are stored in the Flight Ops and Quarterly Ops partitions. Flight Ops runs PDQ and TAD to furnish pointing tweaks and target tables, which are uplinked to the spacecraft and shared with the other clusters as needed. The original data are “snapshotted” (i.e., copied on write) to Monthly Ops from Quarterly Ops to support the first pass processing on each monthly dataset, whose results are used to set parameters for the archive processing on Quarterly Ops. The results from Quarterly Ops are snapshotted to the Archive cluster for export formatting and writing to disc, freeing up Quarterly Ops to continue processing other quarters during reprocessing activities. TEST and LAB are test clusters used for algorithm and code development and for running science analyses, respectively. The SOC operations cluster architecture is shown in Figure 2.10.

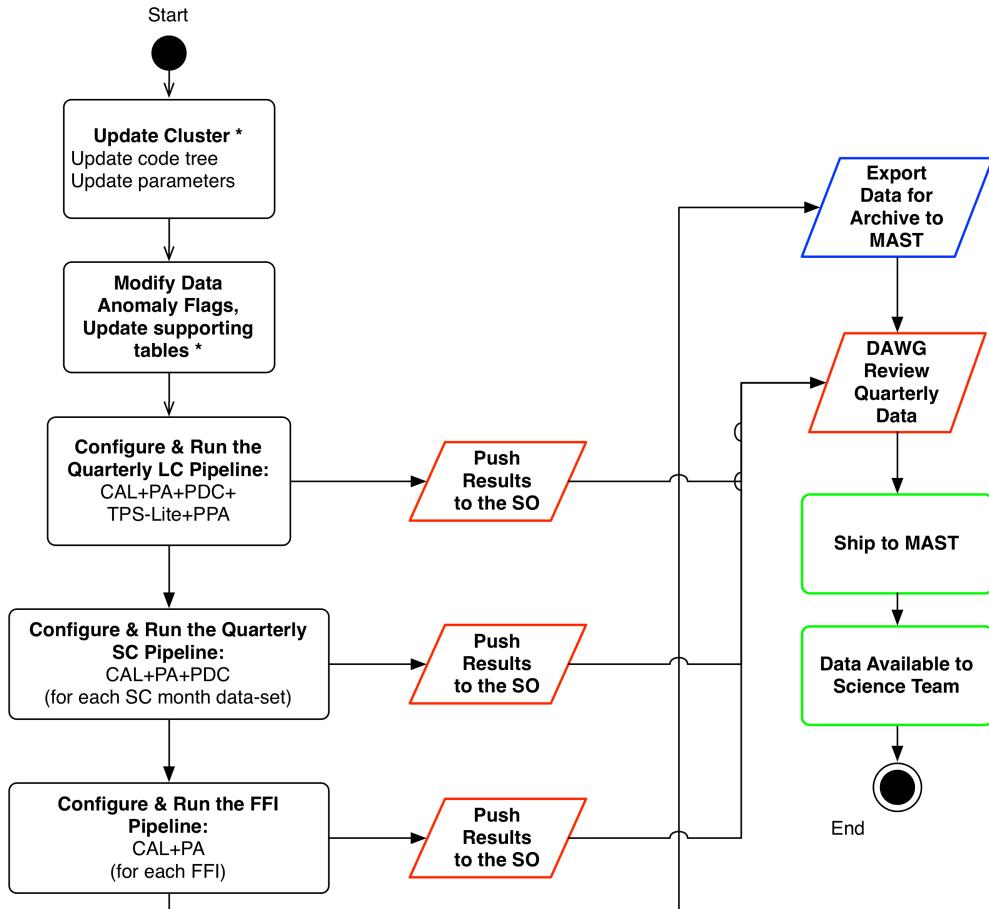
Figure 2.11 shows the operational workflow for processing each quarterly dataset to generate the calibrated pixel and light curve products. The operational environment is updated as necessary to accommodate modifications in code, software parameters, and/or data anomaly flags, which specify which cadences to ignore in processing and for what reason. For this scenario, prelimi-



**Figure 2.10** The SOC operations cluster architecture.

nary processing of the data has already been performed and the results reviewed to identify updates to anomaly flags, etc. Next, triggers are updated and the LC pipeline is kicked off to calibrate the pixels, extract the photometry, correct the light curves for systematic errors, monitor the performance of the instrument and characterize the photometric precision of each light curve. The triggers are updated and the SC pipeline is kicked off. After the SC processing is completed, the triggers are updated and the FFI pipeline is fired. As the pipelines complete and return their data to the Data Store, the data are made available to the Science Office and the Data Analysis Working Group (DAWG) for the purpose of monitoring the performance of the instrument and the pipeline, and for assessing the quality of the resulting data products to inform the generation of Data Release notes accompanying the archival data products.

Figure 2.12 shows the operational workflow for searching multiple quarters of data for transit-like features and fitting models and constructing diagnostic tests for each one found. The operational environment is updated as necessary to accommodate modifications in code, software parameters, and supporting tables. The triggers are updated and the TPS+DV pipeline is fired. Once the processing is complete, the data are made available to the Science Office and the Data Analysis Working Group (DAWG) to assess the performance of the planet search pipeline and to facilitate the identification of KOIs. The results of the planet search were delivered to the Science Analysis System (SAS) during the primary mission for access by the Science Team. In the extended mission, however, the SAS was decommissioned and the planet search results have been archived to the NASA Exoplanet Archive hosted by the NASA Exoplanet Science Institute (NExSci) for access by everyone.

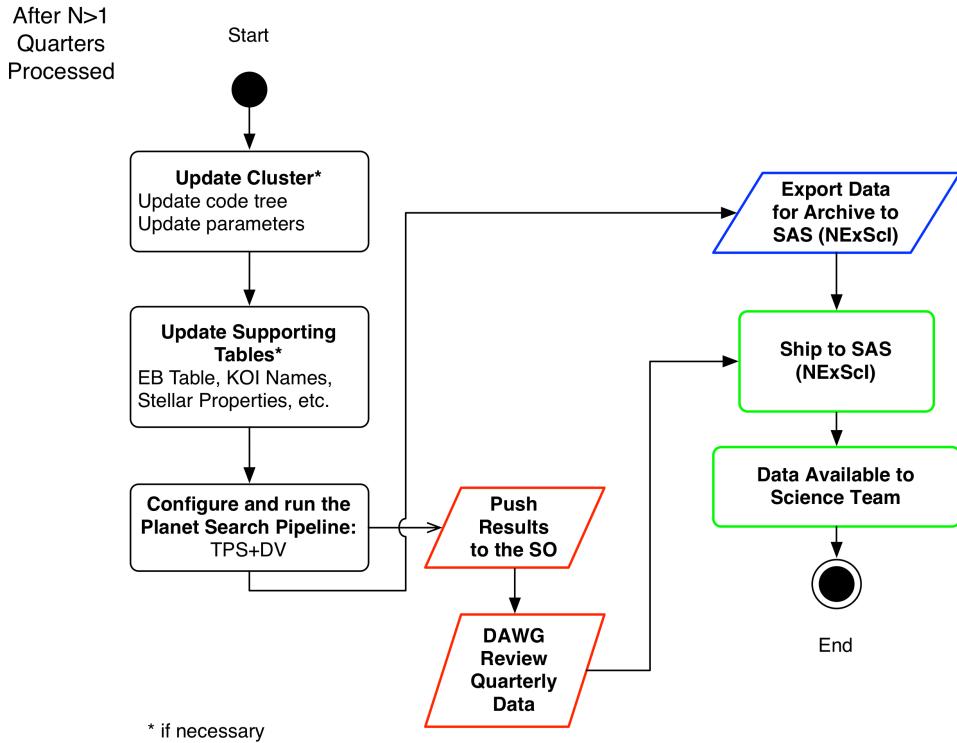


\* if necessary

**Figure 2.11** The quarterly operations workflow for *Kepler* for a single quarter. The purpose of this processing flow is to generate the calibrated pixel and light curve products for a single quarter of data.

## 2.5 Conclusions

The *Kepler* Mission led to the discovery of over 2300 confirmed or validated planets, another 2400 planet candidates, and over 2100 eclipsing binaries. As stunning as the exoplanet discoveries have been, the richness of the astrophysics enabled by *Kepler* is just as breathtaking and enduring. From the detection of asteroseismic pressure mode oscillations in over 15,000 stars, the detection of mixed gravity and pressure modes in red giants, the study of classical variable stars such as Gamma Doradus stars and Delta Scuti stars and hybrids, to novel astrophysics such as highly eccentric “heartbeat” stars and relativistic boosting (van Kerkwijk et al., 2010), the science has been extremely diverse.



**Figure 2.12** The multiple quarter operations workflow for *Kepler*. The purpose of this processing flow is to search for planets and generate diagnostics and limb-darkened model fits for each transit-like feature identified in the search.

The *Kepler* SOC has played a principal role in generating the science data enabling these high impact science results. Developing the SOC was fraught with technical challenges both in terms of dealing with the data volume and in terms of the image and signal processing algorithms implemented in the pipeline. These challenges were met with a flexible pipeline infrastructure and transactional database along with important algorithmic innovations, such as the multi-scale MAP approach to identifying and correcting instrumental systematic errors while retaining intrinsic astrophysical signals to the greatest degree possible. The transiting planet search components of the pipeline saw important innovations, too, enabling exoplanetary science results, including the implementation of an over-complete wavelet transform-based adaptive matched filter that was coupled with  $\chi^2$  statistical vetoes to increase the discriminatory power against instrumental and non-exoplanetary transients in the flux time series data. The success of *Kepler* and the SOC has spurred other missions such as ESA's PLATO Mission and NASA's Transiting Exoplanet Survey Satellite (TESS) Mission. In fact, the *Kepler* SOC is being retooled for use on the TESS Mission to generate the light curves and search for Earth's nearest neighbors starting in 2018 (Ricker et al., 2015).

## Bibliography

- Akeson, R. L., Chen, X., Ciardi, D., et al., 2013. “The NASA Exoplanet Archive: Data and Tools for Exoplanet Research,” PASP, 125, 989
- Allen, C., Klaus, T., & Jenkins, J. 2010. “Kepler Mission’s Focal Plane Characterization Models Implementation,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401E–77401E–8
- Anglada-Escudé, G., Amado, P. J., Barnes, J., et al., 2016. “A Terrestrial Planet Candidate in a Temperate Orbit around Proxima Centauri,” Nature, 536, 437
- Borucki, W. J., Koch, D., Basri, G., et al., 2010. “Kepler Planet-Detection Mission: Introduction and First Results,” Science, 327, 977
- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A., 2011. “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” AJ, 142, 112
- Bryson, S. T., Jenkins, J. M., Peters, D. J., et al. 2010a. “The Kepler End-to-End Model: Creating High-Fidelity Simulations to Test Kepler Ground Processing,” in Proc. SPIE, Vol. 7738, Modeling, Systems Engineering, and Project Management for Astronomy IV, 773808
- Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al., 2010. “The Kepler Pixel Response Function,” ApJL, 713, L97
- Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010c. “Selecting Pixels for Kepler Downlink,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401D
- Caldwell, D. A., van Cleve, J. E., Jenkins, J. M., et al. 2010. “Kepler Instrument Performance: An In-Flight Update,” in Proc. SPIE, Vol. 7731, in Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave, 773117
- Chandrasekaran, H., Jenkins, J. M., Li, J., et al. 2010. “Semi-Weekly Monitoring of the Performance and Attitude of Kepler Using a Sparse Set of Targets,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401B
- Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M., 2000. “Detection of Planetary Transits Across a Sun-like Star,” ApJL, 529, L45
- Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al., 2012. “The Derivation, Properties, and Value of Kepler’s Combined Differential Photometric Precision,” PASP, 124, 1279

- Coughlin, J. L., Thompson, S. E., Bryson, S. T., et al., 2014. “Contamination in the Kepler Field. Identification of 685 KOIs as False Positives via Ephemeris Matching Based on Q1-Q12 Data,” AJ, 147, 119
- Gautier, III, T. N., Charbonneau, D., Rowe, J. F., et al., 2012. “Kepler-20: A Sun-like Star with Three Sub-Neptune Exoplanets and Two Earth-size Candidates,” ApJ, 749, 15
- Gilliland, R. L., Chaplin, W. J., Jenkins, J. M., Ramsey, L. W., & Smith, J. C., 2015. “Kepler Mission Stellar and Instrument Noise Properties Revisited,” AJ, 150, 133
- Gilliland, R. L., Chaplin, W. J., Dunham, E. W., et al., 2011. “Kepler Mission Stellar and Instrument Noise Properties,” ApJS, 197, 6
- Haas, M. R., Batalha, N. M., Bryson, S. T., et al., 2010. “Kepler Science Operations,” ApJL, 713, L115
- Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S., 2000. “A Transiting “51 Peg-like” Planet,” ApJL, 529, L41
- Huber, D., 2016. “Precision Stellar Astrophysics in the Kepler Era,” ArXiv e-prints, arXiv:1604.07442
- Huber, D., Silva Aguirre, V., Matthews, J. M., et al., 2014. “Revised Stellar Properties of Kepler Targets for the Quarter 1-16 Transit Detection Run,” ApJS, 211, 2
- Jenkins, J. M., 2002. “The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets,” ApJ, 575, 493
- Jenkins, J. M., & Dunnuck, J. 2011. “The Little Photometer that Could: Technical Challenges and Science Results from the Kepler Mission,” in Proc. SPIE, Vol. 8146, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 814602
- Jenkins, J. M., Peters, D. J., & Murphy, D. W. 2004. “An Efficient End-to-End Model for the Kepler Photometer,” in Proc. SPIE, Vol. 5497, Modeling and Systems Engineering for Astronomy, ed. S. C. Craig & M. J. Cullum, 202–212
- Jenkins, J. M., Smith, J. C., Tenenbaum, P., Twicken, J. D., & Van Cleve, J. 2012. “Planet Detection: The Kepler Mission,” in Advances in Machine Learning and Data Mining for Astronomy, ed. M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (Chapman and Hall, CRC Press), 355–381
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010. “Transiting Planet Search in the Kepler Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D

- Klaus, T. C., McCauliff, S., Cote, M. T., et al. 2010a. “Kepler Science Operations Center Pipeline Framework,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774017
- Klaus, T. C., Cote, M. T., McCauliff, S., et al. 2010b. “The Kepler Science Operations Center Pipeline Framework Extensions,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774018
- Lépine, S., Hilton, E. J., Mann, A. W., et al., 2013. “A Spectroscopic Catalog of the Brightest ( $J < 9$ ) M Dwarfs in the Northern Sky,” AJ, 145, 102
- Li, J., Allen, C., Bryson, S. T., et al. 2010. “Photometer Performance Assessment in Kepler Science Data Processing,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401T
- Mayor, M., & Queloz, D., 1995. “A Jupiter-Mass Companion to a Solar-Type Star,” Nature, 378, 355
- McCauliff, S., Cote, M. T., Girouard, F. R., et al. 2010. “The Kepler DB: A Database Management System for Arrays, Sparse Arrays, and Binary Data,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400M
- Meibom, S., Barnes, S. A., Latham, D. W., et al., 2011. “The Kepler Cluster Study: Stellar Rotation in NGC 6811,” ApJL, 733, L9
- Mullally, F., Coughlin, J. L., Thompson, S. E., et al., 2015. “Planetary Candidates Observed by Kepler. VI. Planet Sample from Q1–Q16 (47 Months),” ApJS, 217, 31
- Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010. “Pixel-Level Calibration in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401X
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al., 2015. “Transiting Exoplanet Survey Satellite (TESS),” Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003
- Smith, J. C., Morris, R. L., Jenkins, J. M., et al., 2016. “Finding Optimal Apertures in Kepler Data,” PASP, 128, 124501
- Smith, J. C., Stumpe, M. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning II – A Bayesian Approach to Systematic Error Correction,” PASP, 124, 1000
- Stumpe, M. C., Smith, J. C., Catanzarite, J. H., et al., 2014. “Multiscale Systematic Error Correction via Wavelet-Based Bandsplitting in Kepler Data,” PASP, 126, 100

- Tenenbaum, P., & Jenkins, J. M. 2010. “Focal Plane Geometry Characterization of the Kepler Mission,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401C
- van Kerkwijk, M. H., Rappaport, S. A., Breton, R. P., et al., 2010. “Observations of Doppler Boosting in Kepler Light Curves,” ApJ, 715, 51
- Wolfgang, A., Rogers, L. A., & Ford, E. B., 2016. “Probabilistic Mass-Radius Relationship for Sub-Neptune-Sized Planets,” ApJ, 825, 19



## CHAPTER 3

---

# TARGET AND APERTURE DEFINITIONS: SELECTING PIXELS FOR KEPLER DOWNLINK

STEPHEN T. BRYSON<sup>1</sup>, JON M. JENKINS<sup>1</sup>, TODD C. KLAUS<sup>2</sup>, MILES T. COTE<sup>1</sup>, ELISA V. QUINTANA<sup>3</sup>, JENNIFER R. CAMPBELL<sup>4</sup>, KHADEEJAH ZAMUDIO<sup>4</sup>, HEMA CHANDRASEKARAN<sup>2</sup>, DOUGLAS A. CALDWELL<sup>2</sup>, JEFFREY E. VAN CLEVE<sup>2</sup>, AND MICHAEL R. HAAS<sup>1</sup>

<sup>1</sup>NASA Ames Research Center, Moffett Field, CA USA 94035, <sup>2</sup>Stinger Ghaffarian Technologies, Inc./NASA Ames Research Center, Moffett Field, CA USA 94035 <sup>3</sup>The SETI Institute/NASA Ames Research Center, Moffett Field, CA USA 94035, <sup>4</sup>Wyle Labs/NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** The *Kepler* Mission monitors  $\sim$ 165,000 stellar targets using 42 2200  $\times$  1024 pixel CCDs. Onboard storage and bandwidth constraints preclude the storage and downlink of all 96 million pixels per 30-minute cadence, so the *Kepler* spacecraft downlinks a specified collection of pixels for each target. These pixels are selected by considering the object brightness, background, and the signal-to-noise ratio (SNR) in each pixel, and maximizing the signal-to-noise ratio of the target. This paper describes pixel selection, creation of 772 spacecraft apertures that efficiently capture selected pixels, and aperture assignment to a target. Engineering apertures, short cadence targets, and custom-specified shapes are discussed. This chapter is largely based on an updated version of Bryson et al. (2010b).

**Keywords:** *Kepler*, exoplanet, transit, pixel selection

### 3.1 Introduction

The *Kepler* Mission nearly continuously observes  $\sim$ 165,000 target stars in *Kepler*'s 116-square-degree Field of View (FOV) seeking to discover Earth-size planets transiting solar-like stars by detecting photometric signatures of transits (Borucki et al., 2010; Koch et al., 2010). Data is collected and stored for monthly downlink, and the data is processed in the Science Operations Center (SOC – Middour et al., 2010; Haas et al., 2010). Only a limited amount of pixel data can be stored onboard for later downlink to the ground, necessitating the creation of target masks to capture “postage stamps” in the focal plane containing the pixels of interest for each the target stars. The Target and Aperture Definitions (TAD) pipeline module identifies the pixels required to extract photometric measurements for each target star, and also creates and assigns masks used by the flight software to choose which pixel data are stored onboard on the solid state recorder (SSR) during science observations.

Every observed target is readout every 6.52 seconds and co-added 270 times into 29.4-minute observations, referred to as Long Cadence (LC) data. A smaller number of targets, at most 512, is co-added 9 times into 58.8 second Short Cadence (SC) observations. These data are collected nearly continuously for about 30 days and downlinked via high-bandwidth Ka-band transmissions. However, onboard storage is not sufficient to store 30 days of 96 million pixels taken at LC (in addition to the SC and engineering data and FFIs). Moreover, bandwidth constraints of the Ka-band transmission limits the amount of data that may be downlinked, making it impossible to downlink all 96 million pixel values that are collected with each 29.4-min LC observation. It is therefore required to identify the pixels that provide the best data for the target stars and select these pixels for storage and downlink.<sup>1</sup> This paper describes the method by which the required pixels for each target are determined. *Kepler* was placed in an Earth-trailing solar orbit on March 6, 2009. Every  $\sim$ 93 days the spacecraft rotates 90 degrees, referred to as a *quarterly roll*, which changes what stars are on which CCD. Every quarter the pixel selection process is repeated with a new collection of targets<sup>2</sup> (Haas et al., 2010).

Figure 3.1 shows the TAD module in the context of the *Kepler* Science Pipeline. The TAD component was used to define target definition tables that were uploaded to the *Kepler* spacecraft, and determine which pixels are downlinked. In addition to target definitions, the algorithms described in this paper determine which pixels are optimal for aperture photometry. After the data is collected and processed, it can be used to refine optimal aperture selection, for example as described below for saturated targets, or using the techniques described in Chapter 7.

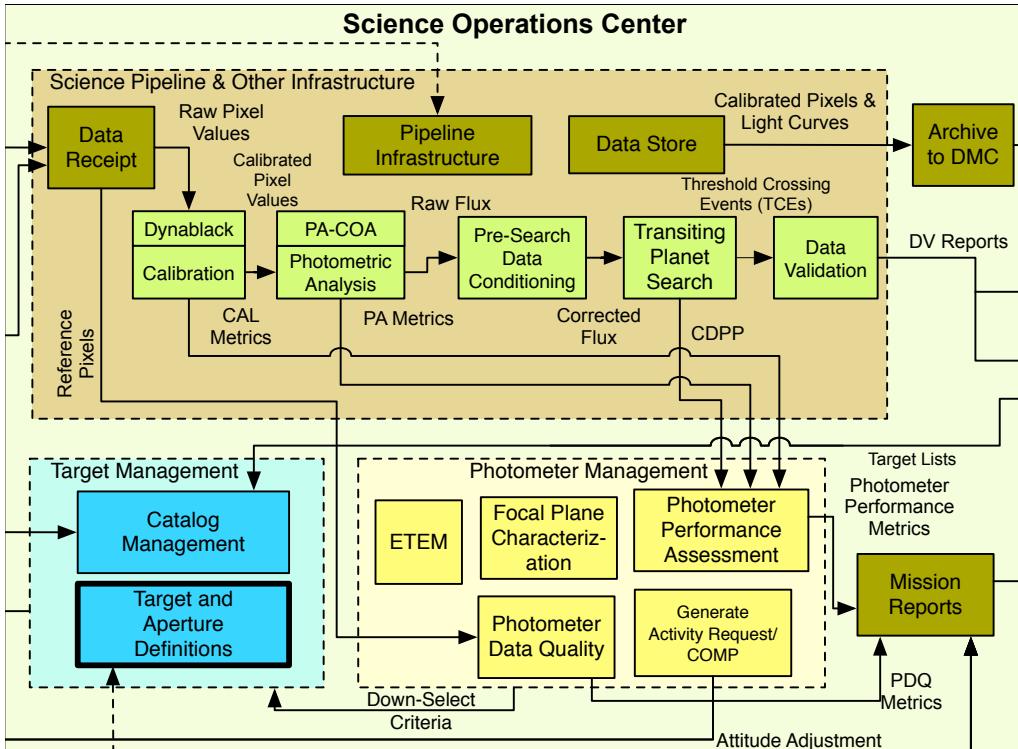
One of the primary goals of pixel selection is to identify pixels that are optimal for aperture photometry, as described in Section 3.2. The high quality of the *Kepler* photometric data (Caldwell et al., 2010) indicates that such pixel selection has been basically successful.

### 3.1.1 The *Kepler* Focal Plane

The *Kepler* focal plane science sensors consist of 42 2200-column-by-1044-row CCDs mounted on 21 electronic modules (Figure 3.2) with an image scale of 3.98" per pixel. The first 20 rows of each CCD are masked with aluminum and are not exposed to the sky in order to provide calibration and diagnostic data as described below. Each CCD is divided into two  $1100 \times 1044$  *output channels*. Each output channel is supplemented by 26 trailing virtual rows, 12 leading serial register columns and 20 trailing virtual columns, giving each output channel  $1132 \times 1070$  addressable pixels. The 12 leading serial register columns and 20

<sup>1</sup>The onboard data compression scheme allows  $\sim$ 65 days of data to be stored, providing robustness against missed Deep Space Network contacts.

<sup>2</sup>Many of the SC targets were changed for each monthly observation, and these target definitions were set up as the same time as for the LC targets.



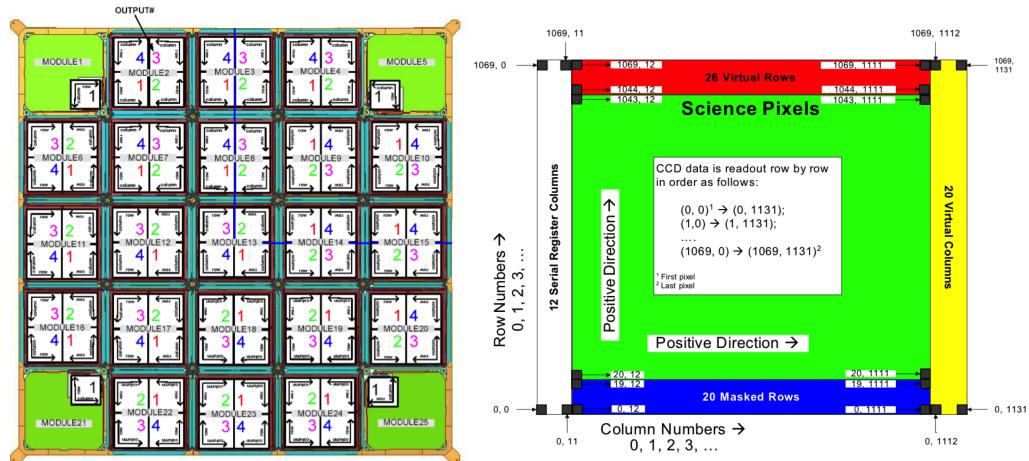
**Figure 3.1** Target and Aperture Definitions (TAD) in the context of the architecture of the SOC. TAD models the sky and the *Kepler* instrument and spacecraft in order to determine which pixels need to be collected to facilitate the ultra high-precision photometry ( $\sim 20$  ppm) required to fulfill *Kepler*'s Mission. TAD was run prior to each quarter's observation using pointing predicts. It was also run after the data were collected using reconstructed pointing information generated by the photometric pipeline in order to provide updated optimal apertures for the photometric measurements.

trailing virtual columns are used to collect black level data for each row. *Kepler*'s lack of a shutter means that pixels are exposed to the sky during readout, which causes image smear along columns. The leading 20 masked and 26 trailing virtual columns measure this smear data. The black level and smear data are called *collateral data* and are used to calibrate the pixel data during ground processing (Quintana et al., 2010). The pair of CCDs on each module provide a contiguous  $2200 \times 2048$  pixel image of a portion of the *Kepler* FOV.

### 3.1.2 Target and Aperture Definitions Task Flow

*Kepler*'s on-board flight system software does not allow specification of arbitrary collections of pixels for each target. A table of 1024 *aperture masks* is loaded onto the spacecraft and the pixels associated with a target are collected with one of these masks. Efficient design of these masks is required to gather the desired pixels without too many excess pixels. This requirement is a primary concern of the work described in this paper.

The pixel selection workflow proceeds along the following lines:



**Figure 3.2** Left: The CCD array on the *Kepler* focal plane, showing the 21 modules, each of which has two  $2200 \times 2048$  pixel CCDs. Each CCD is read out via two output channels. Smaller CCDs used as fine guidance sensors are also shown in each corner, but are not discussed in this paper. Right: The pixel arrangement of each output channel. From Figure 1 of Bryson et al. (2010b).

1. For each target required pixels are identified (Section 3.2).
2. A mask table is created to best match the set of selected pixels for all targets (Subsection 3.3.2).
3. Each target's required pixels are assigned to the smallest mask that contains all required pixels (Subsection 3.3.3).

A single target, particularly a very bright stellar target, may be assigned to more than one mask.

The resulting mask assignment for each target is collected into *target definitions* that include the index in the aperture mask table of the selected aperture mask and the location on the focal plane (module, output, row, column) of that mask.

### 3.1.3 *Kepler* Pixel and Target Types

*Kepler* pixels are collected for several types of targets:

**Stellar Targets:** Point-like sources (Batalha et al., 2010) whose pixels are selected to maximize the SNR (Subsection 3.2.2). Stellar targets are specified by a Kepler ID, which is used to look up pertinent data in the Kepler Input Catalog (KIC – Latham et al., 2005; Brown et al., 2011). Stellar targets may be either LC or SC.

**Custom Targets:** Explicitly specified collections of pixels. Custom targets are defined by a reference pixel position and a set of offsets, one for each pixel, from that reference position. Custom targets are used for non-stellar sources and diagnostic collections of pixels, and may be either SC or LC.

**Background Targets:** Small (nominally  $2 \times 2$ ) sets of pixels that sample the background signal in long cadence. These pixels are selected to support a 2-D polynomial representation of the background (Subsection 3.2.3).

**Reference Pixel (RP) Targets:** Special stellar targets used for managing the science attitude and other diagnostics whose pixels are downlinked bi-weekly via low-bandwidth X-band communications (Chandrasekaran et al., 2010). RP targets are designed to include sufficient information for calibration and analysis (Section 3.4) and are treated separately by the flight system software with their own memory allocation. RP stellar targets share a mask table with LC and SC and also appear on the LC target list. Custom masks are created for RP background targets and the associated collateral data (black and smear).

### 3.1.4 Pixel Selection Requirements

Memory, bandwidth, and flight software design impose several constraints on the final set of pixels selected for downlink:

**Long Cadence (LC):** There may be no more than 170,000 target definitions with no more than 10,000 per output channel and no more than 5.44 million pixels across the FOV. This implies an average of  $\leq 32$  pixels per target definition.

**Short Cadence (SC):** There may be no more than 512 target definitions and no more than 43,520 pixels or an average of  $\leq 85$  pixels per target if 512 targets.

**Background:** There may be no more than 1125 target definitions and no more than 4500 pixels, or an average of 4 pixels per target, for each of the 84 output channels.

**Reference Pixels (RP):** At the beginning of the *Kepler* Mission, there were no more than 96,000 pixels in a set of RP targets across the entire focal plane. As *Kepler* drifted away from Earth, bandwidth degradation reduced the number of RP that could be downlinked in a bi-weekly contact. By the end of the *Kepler* Mission, only 161 stars were collected as reference pixel targets, using 24,277 pixels.

**Aperture Mask Table:** Up to 1024 aperture masks may be defined, using no more than 87,040 pixels. This implies an average of 85 pixels per mask for 1024 masks.

For reasons described in Section 3.4, 252 entries in the aperture mask table are dedicated to supporting RP observations, leaving 772 aperture masks for LC and SC targets.

Basic *Kepler* Mission requirements specify the capability to observe stars with magnitudes between 9 and 15. Simultaneously meeting the above constraints

while providing efficient mask assignment for a wide range of stellar magnitudes is a significant challenge. Observation of stars brighter than magnitude  $Kp = 9$  is allowed with the understanding that tiling efficiency will be significantly degraded, so these observations can use a very large number of pixels.

### 3.2 Pixel Selection

Pixels for stellar targets are selected to maximize the SNR for that target, while background pixels are selected to facilitate the construction of 2-D polynomial representation of the background. Both pixel types are selected based on a model of the signal seen by *Kepler* CCDs created from observed characteristics of the sky (via the KIC and its updates (Huber et al., 2014)) and the *Kepler* photometer, including optical and electronic properties. This model is used to create a synthetic sky image, which is analyzed to determine the desired pixels. The basic strategy for optimal pixel selection is to create two synthetic images: one with all stars in the region of the target star and another with all stars except the target star itself. These images are then used to compare the signal from the target star with the noise from the target star, stellar background, and the instrument.

The techniques in this section are based on techniques developed for creating *Kepler* test data (Jenkins et al., 2004).

#### 3.2.1 Synthetic Image Creation

The synthetic image is created using the following elements:

**Kepler Input Catalog (KIC)**, providing J2000 right ascension (RA,  $\alpha$ ), declination (Dec,  $\delta$ ), and magnitude ( $Kp$ ) in the *Kepler* bandpass of stellar targets in the *Kepler* field.

**Pixel Response Function (PRF) (Bryson et al., 2010a) Model**, an observation-based super-resolution model of how light from a star falls on *Kepler* pixels at different locations in the focal plane. There is one PRF model for each output channel, and this model contains five PRFs that are linearly interpolated to capture intra-channel PRF variations. The observations behind the PRF model were taken at a 15-minute cadence, sufficient to capture the LC behavior of spacecraft pointing jitter. The PRF model includes intra-pixel variability.

**Focal Plane Geometry (FPG) and Pointing Model (Allen et al., 2010)**, which includes measurements of the locations of the CCDs in the *Kepler* focal plane, models of the *Kepler* optics, and models of differential velocity aberrations (DVA). These models are used to determine the pixel location of the central ray of each stellar target. DVA can move a star at the edges of the FOV as much as 0.6 pixels during an observational quarter.

**Saturation Model (Allen et al., 2010)**, which includes information about the well depth of each output channel.

**Zodiacal Light Model**, represented as a mesh of magnitude values on the sky.

**Read Noise Model (Allen et al., 2010)**, observed values of read noise for each output channel.

**Charge Transfer Efficiency (CTE) Model**, which describes how much flux is lost with each charge transfer during readout.

The synthetic image used for pixel selection models the signal in calibrated pixels, so smear due to the lack of a shutter and other instrumental effects is not included. The pixels selected for optimal photometry in Subsection 3.2.2 are used for an entire quarter of  $\sim 93$  days, so for every star the pointing model is used to smear the PRF along the path taken by that star due to DVA during that quarter.

The synthetic image for each output channel is generated star by star. For each star in the KIC that falls on the output channel:

1. The star’s pixel position on the channel is computed from the star’s RA and Dec in the KIC, using the FPG and pointing model (including sub-pixel position).
2. The PRF at the star’s position is evaluated along the path taken by the star due to DVA, creating a smeared PRF covering the entire quarter.
3. The pixels capturing the star’s PRF are normalized to sum to the flux of the star derived from the *Kepler* magnitude in the KIC. The resulting *target-only image* is saved for each target for use in Subsection 3.2.2.
4. The normalized pixels are added to the synthetic image at the appropriate pixel location.

At this point the synthetic image represents the collection of stars in the KIC as they would appear on the CCD without any saturation or zodiacal light background. A copy of the synthetic image, called the *background image*, is made for use in the optimal pixel selection described in Subsection 3.2.2. The zodiacal light is interpolated onto each pixel and added to the synthetic image. Saturation is then iteratively spilled along columns by moving flux that exceeds the well depth up and down the column, with the fraction of flux moving up vs. down being controlled by an input parameter. Finally the CTE model is applied. A comparison of the resulting synthetic image and the in-flight image of the same pixels is shown in Figure 3.3.

There are several sources of error in the generation of the synthetic image, which can result in compromised photometrically-optimal apertures:

**PRF Errors:** Each PRF model is computed as an average of several observed stars without consideration of color. The actual PRF of individual stars will

differ because the optical PSF underlying the PRF has minor color dependence. The PRF varies within a channel, with stronger variation near the edge of the FOV, and this variation is only approximately modeled by the linear interpolation included in the PRF model.

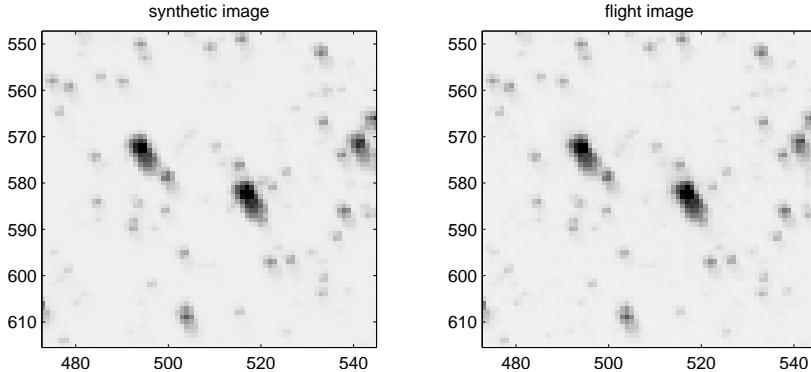
**KIC Errors:** The KIC contains several artifacts that do not correspond to objects in the real sky. Such artifacts can be seen in Figure 3.3 around the bright star at (520, 580), where there is a faint diagonal line of stars extending from the lower left to the upper right through the core of the star. This is likely a diffraction spike misidentified as faint stars, and it does not appear in the flight image. The KIC does not reflect stellar variability, which is unknown for many stars in the *Kepler* field (for stars whose variability is known the brightest magnitude is used). An example of such magnitude errors can be seen in the star at (540, 570) of Figure 3.3, which is brighter in the synthetic image than in the flight image. When a KIC underestimate of a target’s magnitude has been identified, there is a mechanism to provide corrected magnitudes to the pixel selection process. The Stellar Properties Working Group (SPWG) updated the stellar properties in the KIC based on observations of *Kepler* target stars and planetary hosts conducted by the *Kepler* Follow up Observation Program (Huber et al., 2014).

**Saturation Model:** The current saturation model provides the well depth per output channel, assuming that the well depth does not vary within the channel and that the spill up and down a column is symmetric. Both of these assumptions are strongly violated, with variations in both well depth and symmetry. These variations have a higher spatial frequency than could be measured with the bright stars in the *Kepler* FOV. To provide margin against this uncertainty in saturation spill asymmetry, the simulated saturation is extended by 50% in both directions. An approach of explicitly providing the saturation spill per bright star based on observation is described in Subsubsection 3.2.1.1.

**Changing Focus:** The focus of the *Kepler* photometer has been observed to undergo seasonal changes as the Sun angle varies throughout the *Kepler* orbit. In addition, smaller high-frequency focus changes have been observed that are highly correlated with various heaters operating on the spacecraft. These focus changes create PRF changes not captured in the current static synthetic image. Focus changes also induce plate scale variations, so the stars are not placed in the correct positions in the synthetic images.

Because some of these errors were anticipated, the mask assignment process places a one-to-four-pixel halo around the optimal aperture described in Subsection 3.2.2 prior to aperture mask fitting, as described in Section 3.3.

**3.2.1.1 Saturated Targets** During early *Kepler* operations, properly capturing saturated targets proved challenging, due primarily to incorrect *Kepler* magnitudes



**Figure 3.3** A comparison of a synthetic image (left) with the actual flight image of the same region of the sky (right), near the edge of the *Kepler* where the PRF is large. For purposes of comparison with the sky image, the synthetic image is computed for a short time interval for the time of the flight image so that the PRF are not smeared. From Figure 2 of Bryson et al. (2010b).

in the KIC particularly for variable stars and asymmetric saturation spill. In most cases, the 50% buffer added to the simulated saturation described in Subsection 3.2.1 was sufficient so that saturated pixels required for good photometry were captured. In several cases, however, the saturated star's mask was not sufficiently large to fully capture that star's saturated pixels, seriously compromising that star's photometry. Prior to quarter 10 (Q10) operations, manual inspection revealed cases where saturated pixels were not captured. When a lack of full capture was discovered, the missed saturated pixels were recorded and the image creation algorithm in Subsection 3.2.1 was modified to add these saturated pixels to the star's simulated image.

After two years of *Kepler* operations, the process of using *Kepler* data to determine which pixels need to be included in a saturated target's optimal aperture was automated through the introduction of a *saturation map*. The saturation map is a list of bright stars, identified by *Kepler* ID, and, for each bright star, a list of saturated columns with the lowest and highest saturated row. The saturation map is created from the flight full frame images (FFIs) that were collected monthly during the first two years of *Kepler* data collection. Because of *Kepler*'s exceptional pointing stability and repeatability, FFIs from each of *Kepler*'s four observational seasons can be combined to provide several snapshots of the entire *Kepler* focal plane. In this way, observed saturation from every star in the *Kepler* field can be incorporated into the saturation map.

The saturation map is constructed as follows: For each observing orientation, all FFIs taken at that orientation are individually analyzed to create images that show the saturated pixels on the focal plane. These individual saturation images are combined via logical OR operation, resulting in a saturation image per orientation that shows all pixels that were saturated in any FFIs taken at that orientation. Connected sets of saturated pixels in each column are identified in

each orientation’s saturation image. These connected sets of saturated pixels are called *saturation events* and are represented as a column number with lowest and highest row.

The final saturation map is created by comparing saturation events with the KIC, assigning a bright star from the catalog to each saturation event. For very bright stars, several saturation events may be assigned, reflecting the fact that bright stars can saturate several adjacent columns. Each resulting saturation map entry contains a *Kepler* ID and a list of one or more saturation events. The automated assignment of saturation events to KIC stars is manually inspected and corrections are made as necessary. The saturation map adds saturated pixels to the synthetic images used to determine the optimal aperture described in Subsection 3.2.2.

The assignment of saturation events to KIC stars can be ambiguous and even not well defined when there is overlapping saturation from two bright stars in the same column. Such cases are manually resolved by estimating the expected saturation from the stars’ *Kepler* magnitudes. The resulting saturation map entries cannot be expected to be strictly correct in these cases, but overlapping saturation presents a situation in which photometry for the individual contributing star is not possible using saturated pixels. So there is little harm from saturation map errors arising from overlapping saturation from two or more stars.

The saturation map may miss stellar variability due to the once-a-month sampling of the FFIs. The addition of the standard halo described in Subsection 3.3.1 provides a margin against some stellar variability. Visual inspection was continued and the saturation map was augmented when uncaptured saturation was discovered.

When the saturation map is used, the 50% buffer can be removed, resulting in a significant reduction in the pixels captured for saturated targets. This reduction in pixel cost of saturated targets was seen when the saturation map was introduced for Q10.

### 3.2.2 Optimal Pixel Selection

This target-only image created in Subsection 3.2.1 is subtracted from the background image, creating the background image with all stars on this channel except the target. Saturation and CTE are then simulated as described in Subsection 3.2.1. Pixel values  $p_{\text{target}}$  in the target-only image define the signal of each pixel, while the pixel values  $p_{\text{back}}$  in the background image provide the background signal. The SNR of each pixel is estimated as

$$SNR_{\text{pixel}} = \frac{p_{\text{target}}}{\sqrt{p_{\text{target}} + p_{\text{back}} + \nu_{\text{read}}^2 + \nu_{\text{quant}}^2}} \quad (3.1)$$

where  $\nu_{\text{read}}$  is the read noise for this channel and  $\nu_{\text{quant}}$  is the quantization noise given by

$$\nu_{\text{quant}} = \sqrt{\frac{n_C}{12}} \left( \frac{w}{2^{n_b-1}} \right), \quad (3.2)$$

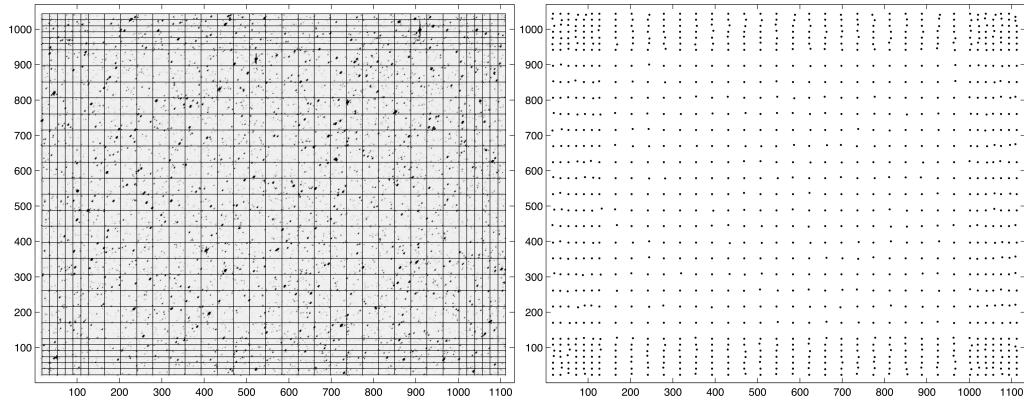
where  $n_C$  is the number of cadences in a co-added observation (270 for LC),  $w$  is the well depth and  $n_b$  is the number of bits in the analog-to-digital converter ( $= 14$ ). This noise formula (Equation 3.1) includes Poisson shot noise of both the target and background pixel values. Given a collection of pixels, the SNR of the collection is given by Equation 3.1 with the individual pixel values  $p_{\text{target}}$  and  $p_{\text{back}}$  replaced by the sum of these pixel values over the pixels in the collection. Optimal pixel selection begins by including the pixel with the highest SNR. The next pixel to be added is the pixel that results in the greatest increase in the SNR of the collection. Initially the collection's SNR will increase as pixels are added. After the bright pixels in the target have been added, dim pixels dominated by noise cause the collection's SNR to decrease. The pixel collection with the highest SNR defines the optimal aperture. Figure 3.4 shows an example of the dependence of the SNR on pixel inclusion.

The background and target images enable crowding to be estimated by calculating the fraction of flux in the optimal aperture that is due to the target star. The resulting *crowding metric* is useful for estimating the dilution of flux from the target in the optimal aperture, which has an impact on the detectability of transits (Batalha et al., 2010). The same measure on a  $21 \times 21$  pixel square provides a *sky crowding metric*, useful for identifying uncrowded stars. The fraction of each target's flux that falls in its optimal aperture is also computed. Both the flux fraction and crowding metric are used in PDC (see Chapter 8) to correct the mean flux level of each light curve for these effects.

### 3.2.3 Background Pixel Selection

Background targets are selected to create a 2-D polynomial representation of the background on each channel. This background polynomial is subtracted from the pixel values prior to aperture photometry (Twicken et al., 2010). To lead to good polynomials the background targets should be homogeneously and uniformly distributed on the output channel. Because the accuracy of background polynomials increases with increasing density of data points but diminishes near the edge of the polynomial domain, more background targets are placed at the edges of the output channel. To achieve this distribution, the intersections of an irregular Cartesian mesh are used for the initial guess at background target positions. Because this mesh is the product of two linear meshes we cannot get a homogeneous distribution that provides exactly 1,125 targets, but a choice of  $31 \times 36 = 1,116$  mesh lines is close.

Because most pixels on an output channel do not contain stars, a pixel in the synthetic image is considered to be background if its value is less than the domi-



**Figure 3.5** Left: The grid used to seed background target positions on an output channel. Right: Final locations of  $2 \times 2$  pixel background targets. From Figure 4 of Bryson et al. (2010b).

nant mode of the pixel histogram. A  $2 \times 2$  pixel background target is considered valid if all four pixels are below the background threshold. Initially, a background target is placed at each mesh intersection. If that target is not a valid background target, increasingly larger boxes centered on the mesh intersection are searched until either a valid background target location is found or the box exceeds a maximum size. In the latter case the target with the smallest summed pixel value is chosen. The performance of this search is enhanced by the creation of a  $2 \times 2$ -average binned image, so that the search in each box is done by taking the minimum of the binned image in that box. The initial mesh and resulting background locations for one example are shown in Figure 3.5.

### 3.3 Mask Creation and Assignment

The pixel selection process in Section 3.2 combined with custom apertures produces several thousand uniquely-shaped pixel apertures across the *Kepler* FOV. These shapes must be mapped to the 772 aperture masks loaded onto the space-craft in a way that minimizes the number of excess pixels collected. This presents a difficult combinatorial problem, for which an optimal solution was not supported by available resources during *Kepler* development. A simpler, near-optimal approach was developed that performs an iterative statistical estimate to determine which pixel apertures may be best used as entries in the mask table.

Our approach is tailored to produce aperture masks efficient for dimmer targets, where the vast majority of pixels are found, and relatively poor for bright targets. This inefficiency for high-priority bright targets is addressed by creating several manually-designed masks.

The mask creation and assignment process proceeds in three stages:

1. Required pixel apertures are defined by adding pixel margin to the requested pixels (Subsection 3.3.1).

2. The mask table is created from an analysis of the resulting final set of pixel apertures, supplemented by manual mask creation for bright targets (Sub-section 3.3.2).
3. The required pixels for each target are mapped to one or more aperture masks in the mask table, creating the final target definitions (Subsection 3.3.3).

The mask table and target definitions are then delivered for quarterly upload to the spacecraft.

### 3.3.1 Required Pixels: Adding Pixel Margin

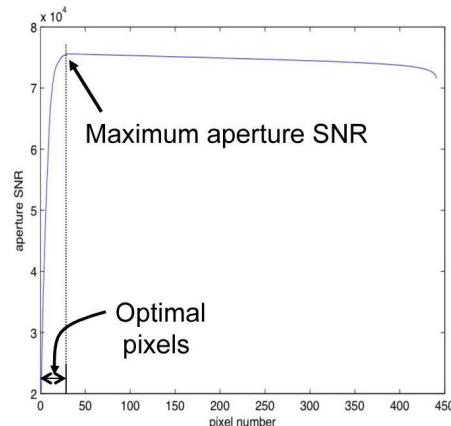
Once the pixels for a target have been specified, either by the optimal pixel selection process in Section 3.2 for stellar targets or by explicit specification for custom targets, the target may be assigned extra pixels. These extra pixels are of two types:

**Halos:** Rings of pixels around the specified pixels for a target, typically applied as a margin against uncertainty. A pixel is added to the halo if any of the eight adjacent pixels, including corners, is included in the target's specified pixels. As many as four halos may be added, which are added iteratively. For example, the third halo treats the previous two halos as specified pixels for the target.

**Undershoot Column:** An extra column of pixels to the left (“upstream” in the pixel readout) of the specified pixels that provides data for the undershoot correction algorithm (Quintana et al., 2010, see also Chapter 5).

Halos and the undershoot column can be specified on a target-by-target basis, and when both are present the halos are applied first. The default for stellar targets is to add one halo and undershoot column, and the default for custom targets is no halo or undershoot column.

The final pixels, including the specified pixels and any halo or undershoot column pixels, are called *required pixels*.



**Figure 3.4** The aperture SNR curve built up as pixels are added in order of decreasing pixel SNR. The optimal set of pixels is defined as the set of pixels required to reach the maximum of this curve. From Figure 3 of Bryson et al. (2010b).

### 3.3.2 Mask Table Creation

We break the table of 1,024 aperture masks into four sections:

**Dim Target Masks:**  $N_{\text{dim}}$  masks that are algorithmically generated by an iterative statistical analysis of the required targets, creating masks that fit required apertures with the smallest number of excess pixels.

**Dedicated Masks:**  $N_{\text{dedicated}}$  masks that are automatically set equal to the required pixels for specific targets. These masks are used for oddly shaped diagnostic targets or very bright high-value stars that are difficult to fit efficiently using other masks.

**Bright Target Masks:**  $N_{\text{bright}}$  masks that are specially generated to fit long saturated columns or large cores of bright stars in order to improve the efficiency of non-dedicated masks on bright stars.

**Reference Pixel Target Set Masks:**  $N_{\text{RPTS}} = 252$  masks that are specifically designed to collect collateral data (black level and smear) for reference pixel targets, as described in Section 3.4.

The dim target portion is the largest section of the mask table (currently  $N_{\text{dim}} = 736$ ) and is filled in using the following algorithm:

1. The aperture mask table is initialized to contain simple geometric shapes.
2. Mask assignment is performed using the full target set as described in Sub-section 3.3.3.
3. The unique required apertures are identified.
4. Masks that are perfect fits to some required apertures are identified and set aside.
5. Masks that were not perfect fits are sorted in descending order of the total number of excess pixels associated with that mask.
6. The masks in the last step are replaced with shapes from the unique set of required apertures for targets dimmer than a specified magnitude until the dim target region of the mask table is filled.

This process is iterated twice in nominal use. The result is the dim target portion of the mask table containing masks appropriate for dim stars, which make up the bulk of the target set.

The bright target portion of the mask table (currently  $N_{\text{bright}} = 28$ ) is filled in using two methods. First, the magnitude range between 7 and 11 is divided into several magnitude intervals (for the current mask table, 18 intervals are used). Within each interval the required pixels of all targets with *Kepler* magnitudes within that interval are combined to make a single aperture mask that fits all

targets within that interval. For targets brighter than magnitude 7, masks are hand-specified to capture their saturation and large cores.

The dedicated mask portion of the mask table is empty prior to final mask assignment and is filled in as described in Subsection 3.3.3.

### 3.3.3 Mask Assignment

Once the mask table is complete, these masks can be assigned to target definitions. Because SC targets are also on the LC target list, only the LC list needs to be considered. RP targets are assigned masks separately, as described in Section 3.4.

The majority of target definitions are assigned aperture masks by choosing the mask that contains the fewest pixels and that completely includes the target's required pixels. Such a mask and the position of the target definition inside the mask are chosen by convolving the target's required pixels (rotated by 180°) with the aperture mask pixels. If the convolution contains values that are equal to the number of required pixels, then the target's pixels fit in the mask. The position of the mask relative to the target pixels is determined by the centroid of all entries in the convolution with values equal to the number of required pixels.

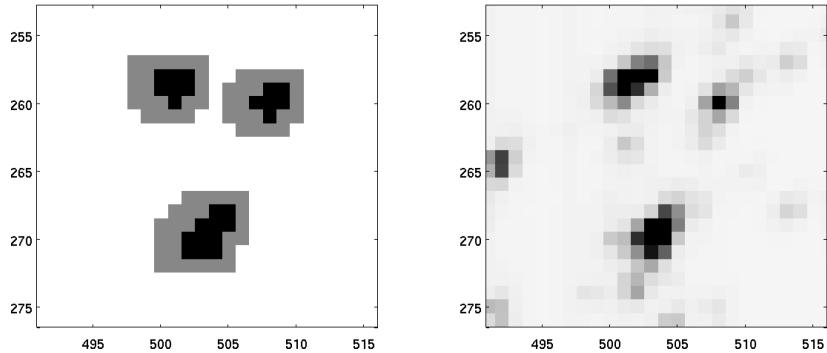
If no aperture mask contains the required pixels, the required pixel set is divided by two along the shorter axis and a new attempt is made to find an aperture mask for each piece. This division approach is applied iteratively until masks are found that cover all required pixels.

Dedicated mask targets are not algorithmically assigned masks by the above process. If there is not already an aperture mask that exactly fits the target's required pixels (several targets may use the same dedicated mask), a slot in the dedicated mask region of the aperture mask table (see Subsection 3.3.2) is filled in by the required pixels of this target. This process assumes that the dedicated mask portion of the mask table has been sized to accommodate all dedicated mask targets.

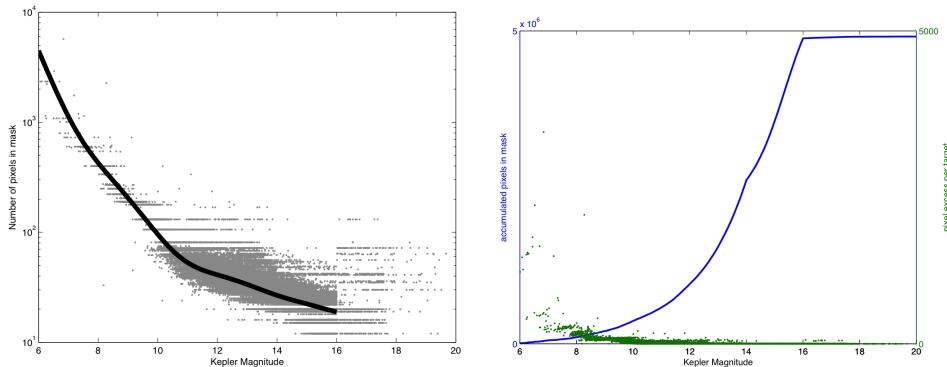
Examples of targets and their assigned masks are shown in Figure 3.6. In operations the number of unused pixels downlinked due to masks over-fitting targets is 4%, mostly from bright stars. Figure 3.7 shows the number of mask pixels per target, the cumulative pixel count and the number of excess pixels per target as functions of target magnitude. We see that the excess pixels are mainly in the bright targets while the dimmer targets, which make up the overwhelming majority of targets, are well fit.

## 3.4 Reference Pixel Targets

A primary design consideration for reference pixels is the very limited number of pixels permitted in an X-band transmission due to low bandwidth. The 96,000 available pixels are used to provide attitude determination and diagnostics on 84 output modules (Chandrasekaran et al., 2010). The requirement to calibrate



**Figure 3.6** Left: Examples of the masks (gray) selected once the halo and undershoot column are added to the optimal aperture (black). Right: A flight image showing the targets captured by the masks from left. From Figure 5 of Bryson et al. (2010b).



**Figure 3.7** Left: The number of pixels per mask vs. target magnitude for stellar targets including pixel halo and undershoot column. The pixel number for each mask is shown by a gray dot, and a 10th order robust polynomial fit showing the mode of the pixel number distribution is shown by the black line. The smallest mask (for a 1-pixel optimal aperture) is  $4 \times 3$ . Banding for the brighter stars is due to the smaller number of masks available for these stars. Right: Cumulative number of pixels in a target's mask vs. magnitude (solid line, left axis), showing that most pixels are on dim targets, and the number of excess pixels per target (dots, right axis), showing that bright targets have the largest pixel excess and dim targets have very little pixel excess. From Figure 6 of Bryson et al. (2010b).

stellar targets for this purpose implies that black-level and smear data for each target must be included in the 96,000 pixel budget. Late in the mission this budget became much smaller as the bandwidth decreased due to the increasing distance of the spacecraft from the Earth.

The target distribution strategy for reference pixels is to assign up to five bright, unsaturated targets to each of the 84 output channels. Channels near the edge of the FOV are preferentially given five targets to increase the attitude determination algorithm's sensitivity to roll. At the beginning of the mission, while bandwidth is high, all channels have at least three reference pixel stellar targets to support plate scale measurements. Special dynamic range targets are also included to provide measurements of the difference between the minimum

and maximum pixel values on each channel, and can quickly indicate if there are problems with the functionality of a given channel. For the first half of the mission, all reference pixel targets are stellar and dynamic range targets. As the bandwidth decreased, smaller rectangular custom targets were used for diagnostics on some output channels starting in quarter 9.

For each stellar target, two halos and an undershoot column are applied, and a stellar target definition is created by assigning a mask using the algorithm described in Subsection 3.3.3. The dynamic range target definitions are created in the same way, but do not require the extra halos or undershoot columns for the mask assignment. Separate target definitions are created to collect background, black, and smear pixels with custom masks that are created to collect the pixels in the most efficient way.

Background pixels are chosen around each stellar target that are close to the local background mode (computed using a synthetic image to compute the local minima) and that are also not corrupted by smear. A custom mask is assigned to the background pixels by assigning the corner pixel of the CCD channel at position (1, 1) to the background target definition and creating a single ‘super mask’ that contains the offsets of all background pixels relative to this target definition pixel. Black and smear pixels that lie in the projection of stellar and background pixels are collected, as are the black pixels needed to calibrate the smear. The values of the black columns and smear rows are user-specified, whereas the black rows and smear columns depend on the location of the stellar targets and background pixels. A target definition is created for each black column (with row index = 1) and for each smear row (column index = 1). The black super mask can then be defined as having row indices equal to the unique rows of the combined stellar, background, and smear rows (with column indices = 1). Likewise, the smear mask consists of column values equal to the stellar and background columns (with row indices = 1).

### 3.5 Conclusions

The algorithms and approaches described in this paper have been successful in specifying the pixels to be downlinked from the *Kepler* spacecraft. These algorithms are used to provide target definition tables for use by the *Kepler* spacecraft. A variety of sometimes conflicting requirements are met, delivering good photometric performance while acquiring a small number of unneeded pixels.

### Bibliography

- Allen, C., Klaus, T., & Jenkins, J. 2010. “Kepler Mission’s Focal Plane Characterization Models Implementation,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401E–77401E–8
- Batalha, N. M., Borucki, W. J., Koch, D. G., et al., 2010. “Selection, Prioritiza-

- tion, and Characteristics of Kepler Target Stars,” ApJL, 713, L109
- Borucki, W. J., Koch, D., Basri, G., et al., 2010. “Kepler Planet-Detection Mission: Introduction and First Results,” Science, 327, 977
- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A., 2011. “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” AJ, 142, 112
- Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al., 2010. “The Kepler Pixel Response Function,” ApJL, 713, L97
- Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010b. “Selecting Pixels for Kepler Downlink,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401D
- Caldwell, D. A., Kolodziejczak, J. J., Van Cleve, J. E., et al., 2010. “Instrument Performance in Kepler’s First Months,” ApJL, 713, L92
- Chandrasekaran, H., Jenkins, J. M., Li, J., et al. 2010. “Semi-Weekly Monitoring of the Performance and Attitude of Kepler Using a Sparse Set of Targets,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401B
- Haas, M. R., Batalha, N. M., Bryson, S. T., et al., 2010. “Kepler Science Operations,” ApJL, 713, L115
- Huber, D., Silva Aguirre, V., Matthews, J. M., et al., 2014. “Revised Stellar Properties of Kepler Targets for the Quarter 1-16 Transit Detection Run,” ApJS, 211, 2
- Jenkins, J. M., Peters, D. J., & Murphy, D. W. 2004. “An Efficient End-to-End Model for the Kepler Photometer,” in Proc. SPIE, Vol. 5497, Modeling and Systems Engineering for Astronomy, ed. S. C. Craig & M. J. Cullum, 202–212
- Koch, D. G., Borucki, W. J., Basri, G., et al., 2010. “Kepler Mission Design, Realized Photometric Performance, and Early Science,” ApJL, 713, L79
- Latham, D. W., Brown, T. M., Monet, D. G., et al. 2005. “The Kepler Input Catalog,” in Bulletin of the American Astronomical Society, Vol. 37, American Astronomical Society Meeting Abstracts, 1340
- Middour, C., Klaus, T. C., Jenkins, J., et al. 2010. “Kepler Science Operations Center Architecture,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401A
- Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010. “Pixel-Level Calibration in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401X

- Twicken, J. D., Clarke, B. D., Bryson, S. T., et al. 2010. “Photometric Analysis in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774023



## PART II

---

# THE KEPLER PHOTOMETRIC PIPELINE



# CHAPTER 4

---

## DYNAMIC BLACK CORRECTION

BRUCE D. CLARKE<sup>1</sup>, JEFFREY J. KOLODZIEJCZAK<sup>2</sup>, DOUGLAS A. CALDWELL<sup>1</sup>, JEFFREY E. VAN CLEVE<sup>1</sup>, JON M. JENKINS<sup>3</sup>, MILES T. COTE<sup>3</sup>, TODD C. KLAUS<sup>4</sup>, VIC S. ARGABRIGHT<sup>5</sup>

<sup>1</sup>The SETI Institute/NASA Ames Research Center, Mountain View, CA 94035, <sup>2</sup>NASA Marshall Space Flight Center, Huntsville, AL 35808, <sup>3</sup>NASA Ames Research Center, Moffett Field, CA 94035, <sup>4</sup>Stinger Ghaffarian Technologies, Inc./NASA Ames Research Center, Moffett Field, CA 94035, <sup>5</sup>Ball Aerospace & Technologies Corp., Boulder, CO 80301

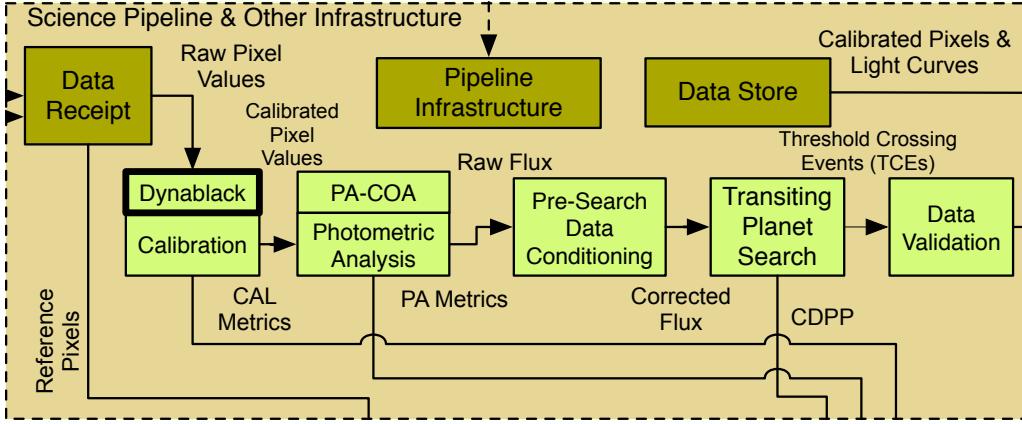
**Abstract.** In order for *Kepler* to achieve its required  $<20$  ppm photometric precision for magnitude 12 and brighter stars, instrument-induced variations in the CCD readout bias pattern (“2-D black image”), which are either fixed or slowly varying in time, must be identified, and the corresponding pixels either corrected or removed from further data processing. The two principle sources of these readout bias variations are: 1) crosstalk between the 84 science CCDs and the four fine guidance sensor (FGS) CCDs, called FGS crosstalk, footnoteWhile there is video crosstalk between the 84 science CCD readout channels, the effects don’t behave as fixed pattern noise, as is the case with the crosstalk from the FGS to the science CCDs, which is caused primarily by the clock driver pulses. and 2) a high-frequency amplifier oscillation on  $<40\%$  of the CCD readout channels. The crosstalk produces a synchronous pattern in the 2-D black image with time-variation observed in  $<10\%$  of individual pixel bias histories. We describe a method of removing the crosstalk signal using continuously-collected data from masked and over-clocked image regions (our “collateral data”) and occasionally-collected full-frame images and reverse-clocked readout signals. We use this same set to detect regions affected by the oscillating amplifiers. The oscillations manifest as time-varying moiré patterns and rolling bands in the affected channels. Because this effect reduces the performance in only a small fraction of the array at any given time, we have developed an approach for flagging suspect data. These flags provide the necessary means to resolve any potential ambiguity between instrument-induced variations and real photometric variations in a target time series. We also evaluate the effectiveness of these techniques using flight data from background and selected target pixels. This chapter is largely an updated version of Kolodziejczak et al. (2010).

**Keywords:** *Kepler*, photometer, CCD, noise, pattern noise, CCD readout, crosstalk, image analysis, photometry

### 4.1 Introduction

The first step of processing *Kepler* data is to perform pixel level calibrations, as shown in Figure 4.1. However, electronic image artifacts identified prior to launch (Van Cleve & Caldwell, 2016) motivated the development of partial

mitigations for these thermally sensitive variations in the bias voltage or “black level” of the CCD measurements. The Dynablack (DYN) module analyzes the black measurements made on the nominal long cadence (LC) data and the FFIs to establish thermal corrections to the black correction coefficients used in the calibration (CAL) module (see Chapter 5). DYN also flags instances of rolling band anomalies, which can inject transit-like features into the science data, causing spurious detections in the transiting planet search (TPS—see Chapter 9). This chapter is an extension of the work presented in Kolodziejczak et al. (2010).



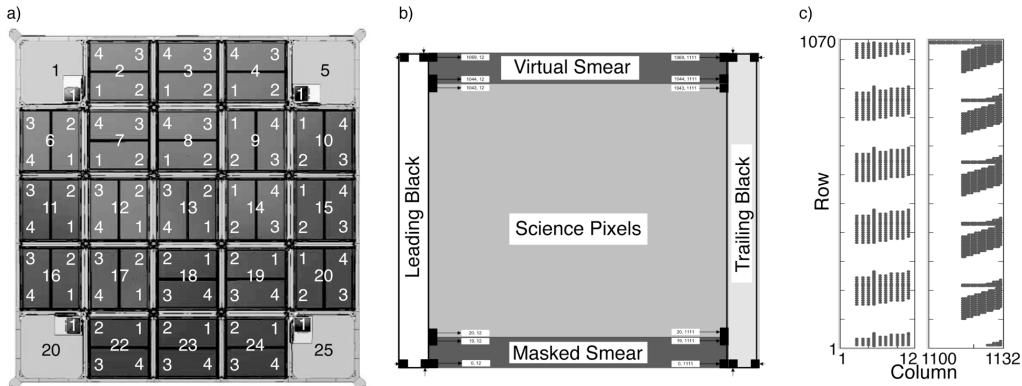
**Figure 4.1** Dynablack (DYN) in the context of the architecture of the SOC. DYN analyses both long cadence (LC) data as well as full frame images (FFIs) to identify improved black level corrections for the calibration (CAL) module. DYN also identifies regions of the CCDs corrupted by rolling band anomalies, which are caused by oscillating operational amplifiers in the analog focal plane electronics.

Within the *Kepler* calibration pipeline module (CAL) there are two types of black correction available; 1) a static 2-D black plus a row dependent 1-D adjustment which is computed cadence by cadence and 2) a dynamic 2-D black which includes per cadence row, column, and crosstalk pixel type dependence. It is the dynamic 2-D black estimate which is described below and is implemented in the standalone *Kepler* Pipeline module DYNABLACK, a sub module of CAL. The process of dynamic 2-D black estimation and the subsequent application of this estimate in CAL together are referred to as Dynamic Black Correction.

## 4.2 Background

The *Kepler* focal plane consists of 84 separate science readout channels (identified as module#.output#) and four fine guidance sensor (FGS) channels, as shown in Figure 4.2a, all of which are read out synchronously. Each channel has several regions available to collect calibration, or “collateral” data (Figure 4.2b). There are two sets of columns of virtual pixels: 1) 12 columns of bias-only pixels resulting from 12 leading pixels in the serial register (“leading black”) and 2) a 20-column serial over-scan region (“trailing black”). There are also two sets of rows of collateral pixels: 1) the first 20 rows, which are covered by an

aluminum mask (“masked smear”), and 2) a 26-row parallel over-scan region (“virtual smear”). During science data collection, a co-added sum of specified columns of the trailing black and rows of both the masked and virtual smear are stored at each cadence for each channel. To enable correction for some of the artifacts, a specific set of artifact removal pixels (ARPs), as illustrated in Figure 4.2c, are collected along with the science data.



**Figure 4.2** a) Focal plane with designations for modules 1–25 and outputs 1–4 for each module. Module.outputs 1.1, 5.1, 20.1 and 25.1 are FGS. b) Identification of collateral regions associated with each output channel. c) Plots indicating the locations of artifact removal pixels (ARPs) in the leading and trailing black regions. These pixels are collected during each long cadence. From Figure 1 of Kolodziejczak et al. (2010).

Science data are available at either short cadence ( $\sim$ 1 minute) for up to 512 targets or long cadence ( $\sim$ 30 minutes) for up to 170,000 targets. All science data are collected with an integration time of 6.02 s with pixels read out at a 3 MHz clock rate. In science collection mode, the full single integration CCD frames are co-added together. Then at the end of the short and long cadence period pre-specified pixels for each target are selected from the co-add, processed, and stored on board. Due to data storage and transmission limitations, only about 6% of the 96 million pixels are stored for eventual transmission to the ground.

*Kepler*'s shutterless operation precludes standard dark frames. Instead, reverse-clocking of the CCDs permits us to measure the bias level throughout the image in the absence of sky signal. Also, a full frame image (FFI) mode permits collection of all the pixels in the focal plane. FFIs are used to examine detector properties, verify pointing, and verify the target aperture definitions. Reverse-clocked data and FFIs were taken periodically throughout the mission (Chandrasekaran et al., 2010).

#### 4.2.1 Description of Image Artifacts

Ground testing prior to flight uncovered several instrumental artifacts, each of which was investigated to understand its cause, impact, and cost to fix or mitigate. These artifacts were extensively characterized on the ground and then

again on-orbit during the commissioning phase of the mission. These investigations determined that several artifacts did not require mitigation. The existing data processing pipeline (see Chapter 5 and Quintana et al. (2010)) or the algorithms described herein handle those with the largest impact, such as those that contribute to the formation of a readout bias pattern. Those relevant to this discussion are briefly described below and illustrated in Figure 4.3.

**FGS Clocking Crosstalk.** Crosstalk from the FGS clock signals to the science CCD video signals injects a complex pattern into the bias image of every science channel with an amplitude up to 20 DN per read (Argabright et al., 2008). Because the FGS and science CCDs share the same master clock, the pattern is spatially fixed; however, the amplitude of the crosstalk is dependent on the temperature of the Local Detector Electronics (LDE). The crosstalk has three distinct components based on the state of the FGS CCDs as the science pixels are read out (see Figure 4.3a): FGS CCD frame transfer, parallel transfer, and serial transfer. Approximately 20% of targets have at least one of the parallel or frame-transfer crosstalk pixels in their aperture. Without mitigation, the crosstalk introduces a small time-varying bias into a target’s flux time series as the LDE temperature changes.

**High-frequency Oscillations.** A temperature-sensitive amplifier oscillation at  $>1$  GHz was detected in some CCD video channels during the artifact investigation. This suggests that the signal may originate from the AD8021 operational amplifiers used extensively in the video signal chain, which may show subtle layout-dependent instability when used at low gains when operated at the upper end of their rated speed range (as is the case for the *Kepler* focal plane electronics). The oscillation’s frequency range, rate of change and pattern among the channels matched closely those characteristics in the dark images, strongly suggesting that the artifact is a moiré pattern (MPD) generated by sampling the high-frequency oscillation at the 3MHz serial pixel clocking rate. Since the characteristic source frequency drifts with time and the temperature of the electronic components by as much as 500 kHz/C, the signal from a given pixel in a series of dark images has a time-varying signature. This signature may be highly correlated with neighboring pixels and yet poorly correlated with slightly more distant pixels. When the oscillation frequency is a harmonic of the serial clocking frequency, a DC shift occurs producing a horizontal band offset from the mean bias-level in the image. As the frequency drifts with temperature, the point on the image where this DC shift occurs moves up or down from sample-to-sample, producing a rolling band artifact (RBA).

Forty-six of the 84 readout channels have never exhibited moiré pattern behavior and an additional nine channels have not exhibited this behavior at a detectable level in flight. While the moiré amplitude per pixel in the remaining channels is significant, its effect on our ability to detect small planets depends on frequency, sum within a target aperture, and variations over time scales of interest to transit detection. The instrument meets the 6-hour precision requirement across the focal plane for the quietest 30% of stars (Jenkins et al., 2010). The

two worst moiré channels, 9.2 and 17.2, exhibit a  $\sim$ 20% increase in 6-hour noise over the focal plane average at 12th magnitude, based on the standard deviation of 6-hour binned flux time series. Such an increase is small compared with the factor of 1.5 spread in the distribution of dwarf star precision at 12th magnitude (Jenkins et al., 2010). Twenty-nine of the 84 CCD channels exhibit moiré pattern noise and/or rolling band artifacts (Caldwell et al., 2010).

For completeness, we briefly mention several other instrumental features in the *Kepler* data which are either accounted for in the baseline calibration scheme (Quintana et al., 2010) or are not currently observed to be variable enough to adversely affect *Kepler* science. The list includes:

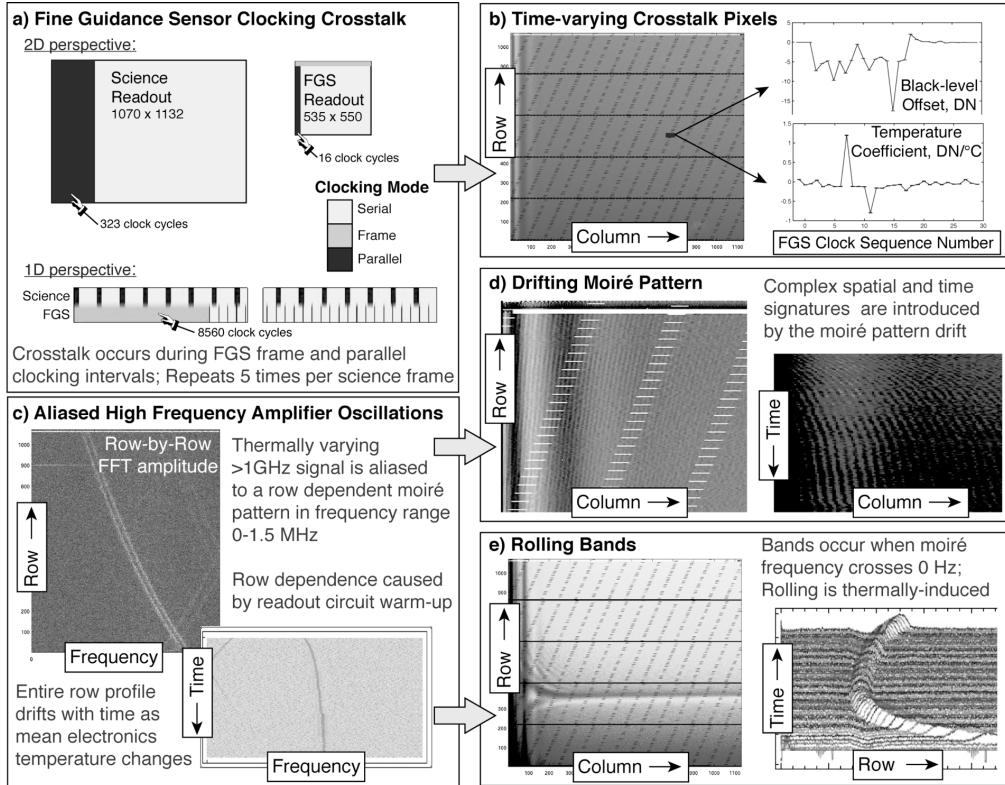
1. Time-varying low-spatial-frequency characteristics, whereby the row-profile of the 2-D black image tends to evolve slowly with time.
2. LDE undershoot, whereby star-like images induce a signal-dependent trailing undershoot in the video output, and
3. Start-of-line ringing (SOLR), whereby a transient signal is initiated at the onset of serial clocking of each row.

The variations at low spatial frequency are corrected using a cadence-by-cadence fit to the trailing black collateral data detailed in the algorithm described in Section 4.3, where we also account for FGS crosstalk and LDE undershoot. The SOLR has been observed to be sufficiently stable over time to avoid photometric precision degradation.

Finally, scene-dependent artifacts arise in two possible ways: 1) as a consequence of the temperature sensitivity of the oscillating LDE component, the thermal transient introduced during readout by the signal from a bright star causes additional localized changes in the detected moiré pattern amplitude and frequency and 2) through bright stars, especially at the ends of saturated column segments, introducing variability in downstream pixels resulting from the LDE undershoot from beyond 20 pixels. How these could be addressed is discussed in Section 4.3.

#### 4.2.2 *Kepler's Noise Floor*

The *Kepler* design places the noise floor for detecting stellar variation at a level that enables detection of Earth-size planets in habitable orbits around 12th magnitude stars. A remotely observed Earth would produce an 84 ppm decrement during a 12-hour solar transit, and to ensure high detection efficiency for similar objects, a  $1\sigma$  noise level of 20 ppm for 6.5-hour time intervals is allocated. Artifact-induced changes in bias level significantly exceeding 20 ppm on this time scale or longer could be confused with the signals resulting from stellar variability. Therefore, the objective of Dynamic Black Correction is to correct bias variations to below the noise floor where possible, and to detect and flag them where correction is not possible. Table 4.1 summarizes implications for the acquired signal levels in electrons and DN.



**Figure 4.3** Illustrative description of pattern noise sources (left) and effects on *Kepler* images (right). a) The FGS pixels are read out synchronously with the science pixels but the difference in size of the two sensors combined with the changes in characteristics of the clocking signals during the parallel and frame transfer intervals produces b) a complex pattern of thermally varying pixels at specific locations across the image. c) An aliased high frequency signal is highly sensitive to LDE component temperature changes due to normal operational warm-up. These produce a smoothly varying frequency change with row. Note that a 0.1% change in the source frequency can produce an aliased frequency change from DC to Nyquist (1.5 MHz) in the readout. One result is d) a moiré pattern with row-dependent frequency (left). The trace of a single row with time shows drift in the frequency (right). The other result is e) bands extending over all columns, which occur when the aliased frequency crosses 0 Hz (left). The time-evolution of a single column shows rolling band in response to temperature changes (right). Plots are meant to show typical qualitative behavior for affected channels, so scales and data sources are suppressed. For reference b) shows flight data from output 20.2, c) shows ground test data from output 14.4, d) shows ground test data from output 9.3, and e) shows ground test data from 13.4. From Figure 2 of Kolodziejczak et al. (2010).

The characteristics of pattern noise that justify the additional attention are that it may be spatially correlated and temporally correlated, whereas ordinary white noise or other broadband types of noise are likely to be less so. For an aperture with 16 pixels, observed for a time interval of 6.5 hours (13 long cadences) noise will grow in the sum at a rate  $\sqrt{16 \cdot 13} \approx 14$  times more slowly than a spatially- and temporally-correlated bias change. On the other hand, pattern characteristics vary widely, so only a small fraction of the integrated observing time field-of-view product is likely to be subjected to a reduction in sensitivity or increased likelihood of false detection. In this paper we use the term “source coverage” to identify the product of time field-of-view.

**Table 4.1** Quantities used to define the level at which artifacts begin to affect *Kepler* science.

	11	12	13
Stellar magnitude (G-type star)			
Detected in-aperture electrons in 6.5 hours, aperture sizes are typical for the magnitude	1.12e+10 in 25 pixel aperture	4.47e+09 in 16 pixel aperture	1.79e+09 in 9 pixel aperture
Earth-equivalent transit decrement, 84 ppm (e-)	939000	376000	150000
Allocated noise including shot noise, 20 ppm for 12th mag. (e-)	168000	89400	51900
Bias change matching allocated noise level in aperture in 6.5 hr (e- read <sup>-1</sup> pixel <sup>-1</sup> )	1.9	1.6	1.6
Bias in DN, nominal 100 e-/DN scale factor (DN read <sup>-1</sup> pixel <sup>-1</sup> )	0.019	0.016	0.017

#### 4.2.3 Acceptable Bias Variations

It is clear from Table 4.1 that space- and time-correlated signals below 0.02 DN/read over time intervals shorter than 6.5 hours would be difficult to distinguish from noise and therefore have little scientific impact. Typically, the acceptable level of bias variation would be constrained by detectability to several  $\sigma$  above the noise level. In this case however, we are able to leverage the similar behavior of pixels acquired at a specific FGS clocking interval or pixels in an extended region of a given readout channel to measure bias variations at several times below the noise level. Since it is no more difficult to observe effects at the noise level than several times above, we simply round the 0.016 DN read<sup>-1</sup> pixel<sup>-1</sup> to 0.02 DN read<sup>-1</sup> pixel<sup>-1</sup>, which is equivalent to 25 ppm of the 6.5-hour signal from a 12th magnitude star. Thresholds for correction and flagging of artifact-induced bias variations are based on limiting changes to 0.02 DN read<sup>-1</sup> pixel<sup>-1</sup>.

#### 4.2.4 Artifact Removal and Flagging Objectives

FGS crosstalk is clearly detectable in data from collateral regions collected during every long cadence. The multiple examples of pixels collected during each FGS parallel and frame clocking interval typically indicate a repeating pattern that shows little or no change over one science readout. It is therefore possible to measure these bias changes on a cadence-by-cadence basis and have high confidence that the science pixels are subject to the same effects. The objective for the FGS algorithms is therefore to remove the FGS crosstalk signal to the level of 0.02 DN read<sup>-1</sup> pixel<sup>-1</sup>.

Collateral data from smear regions also provides a way to measure the amplitude and frequency of moiré patterns on a cadence-by-cadence basis at the beginning and end of each readout interval. Relating this to the row-by-row amplitude and frequency of moiré patterns in the difference between FFIs provides an indirect means of estimating the characteristics of moiré patterns at any readout location and time. Changes in trailing black collateral provide a similar

means to detect and characterize rolling bands. For these, however, confidence in the exact estimate of the artifact-induced bias does not match that of FGS corrections. Such efforts are complicated significantly by complex phase variations and scene-induced changes in readout component temperatures, which lead to short term changes in the oscillating frequencies and local shifts in the moiré patterns. *For these reasons, our objectives with respect to MPD and RBA are to flag regions of the focal plane and times when collateral data indicate that the amplitude of an effect is greater than  $0.02 \text{ DN read}^{-1} \text{ pixel}^{-1}$ .* The impact of a specific moiré pattern amplitude on the peak-to-peak variation in an aperture is reduced by a factor of  $|2 \sin(\pi fn) / (\pi fn)|$ , where  $n$  is aperture width in pixels, and  $f$  is the moiré spatial frequency. This makes it difficult to generalize the severity of the moiré pattern in a flagged region independent of the target apertures. Our approach is to flag data based on amplitude and provide localized severity information, such as frequency, that permit more precise evaluation of the impact on a given target.

The dynamic black correction algorithms as implemented in the DYNABLACK pipeline module include flagging only for RBA. However, DYNABLACK output includes the model descriptions and fit residuals required for MPD analysis and flagging.

### 4.3 Methods

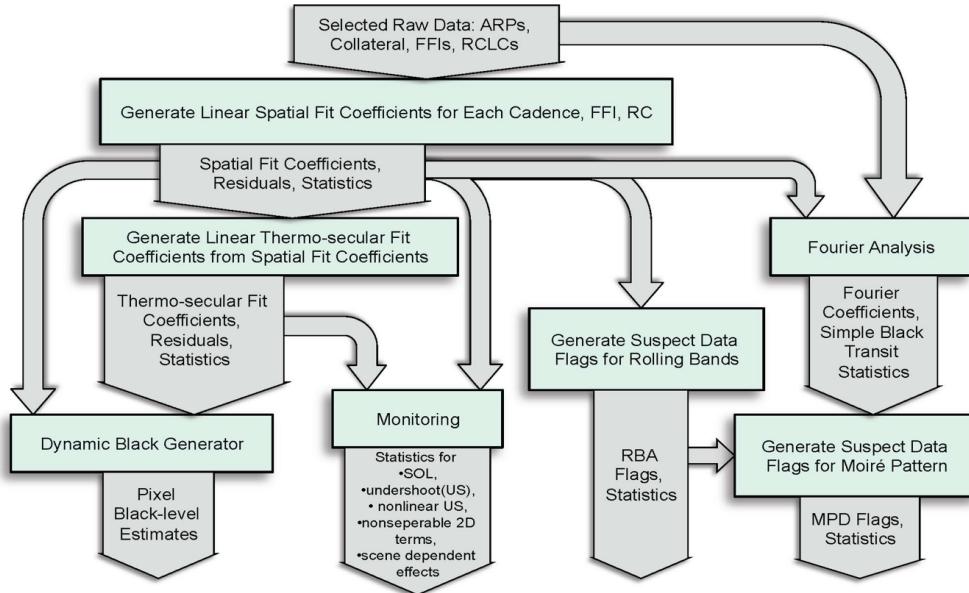
The process developed to mitigate the effects of the pattern noise sources on *Kepler* science includes the elements and data products shown in Figure 4.4. The overall architecture takes advantage of the results of cadence-by-cadence spatial fitting, removing FGS crosstalk and low-spatial frequency variations from fit residuals, which would otherwise complicate detection of the rolling bands and moiré pattern.

The Pipeline module DYNABLACK organizes the dynamic black fit algorithms as follows:

- A1 – Single cadence spatial fits vs. row
- A2 – Single cadence spatial fits vs. column
- B1a – Fit coefficient results from A1 to time and temperature model
- B1b – Fit coefficient results from A2 to time and temperature model

#### 4.3.1 Spatial Fitting

The spatial fitting algorithm is designed to extract information about the time-varying parts of the *Kepler* data stream using pixel and collateral data from each cadence or FFI. We derive the information in the form of fit coefficients and uncertainties based on a model of the observed behavior of each pixel. Let  $X_{RC}$



**Figure 4.4** Architecture diagram showing prototype elements as boxes and data product flow as arrows. The algorithm as implemented in the Pipeline module DYNABLACK does not include the rightmost branch (Fourier Analysis → Generate Suspect Data Flags for Moiré Pattern). From Figure 3 of Kolodziejczak et al. (2010).

represent the raw black level value in DN at row  $R$  and column  $C$ , and let  $Z$  represent the zero offset introduced to prevent negative values. Then we assume,

$$X_{RC} - Z = X_0 + f_{RC} + g_C + W_{RC}^F + W_{RC}^P + U_{RC}, \quad (4.1)$$

where  $X_0$  is a constant,  $f_{RC}$  is the row dependent term, ( $C$  only distinguishes leading and trailing black),  $g_C$  is the column dependent term,  $W_{RC}^F$  is the FGS frame crosstalk-dependent term,  $W_{RC}^P$  is the FGS parallel crosstalk-dependent term, and  $U_{RC}$  is the undershoot dependent part.

Each of these components is defined in Appendix 4-A in terms of linear coefficients. The fit values of all the coefficients, coefficient errors, residuals and associated statistics are the data products of this algorithm.

The fitting process includes initialization and fitting segments. Initialization of FGS clock states as a function of pixel location, selection of scene-dependent exclusion zones and initialization of model components are parts of the initialization segment. Fitting the average of specified cadences and fitting each of a series of single cadences comprise the fitting segment. The FGS frame and parallel crosstalk components occur in repeating patterns defined by the parallel and frame-clocking intervals of the FGS. The term “row” applies to one of 1070 rows of a science channel image, each 1132 pixels long. The actual number of 3MHz clock cycles in a row is 1455, so 323 clock cycles do not result in pixels (science parallel clocking). There are five FGS frames (311,370 clock-cycles) for every science frame, so an overall pattern repeats every 214 rows. In terms of the continuous clock stream, the frame crosstalk signal repeats every 16 pixels for 8,450

clock cycles. It then flat-lines to a constant offset for 109 clock cycles followed by a short 3-clock-cycle waveform. The parallel crosstalk signal repeats every 566 clock cycles with the first 16 cycles having the strongest crosstalk signal followed by another 12–24 cycles that exhibit low-level “ringing”. The remainder of the 566 may be treated as constant except for the last one, which shows variability in some channels (cycles 17–566 are the serial clocking intervals of the FGS frame readout). The parallel pattern begins at FGS clock cycle 8563 with parallel cycle number 3 of 566 and continues *modulo* 566 through the entire FGS clock cycle range to 311,370.

To serve the fitting algorithm we assign the FGS-frame clock sequence numbers 1–16 to the repeating frame pattern, the value 17 to the 109 following pixels and 18–20 to the interval-ending feature. The remainder of the 311370 clock-cycles are assigned the value 0. Similarly, we assign the FGS-parallel clock sequence numbers 1–566 to the repeating parallel pattern starting with value 3 assigned to cycle 8,563, and all preceding cycles are assigned the value 0. For more details on the FGS clocking crosstalk pattern, see Thompson et al. (2016, §2.3.5.1).

After stacking the five FGS frames and chopping off the excess 323 unobserved columns, this provides a pair of mappings from (row, column) to FGS-frame clock sequence number and (row, column) to FGS-parallel clock sequence number. We fit a spatial model in which each pixel’s black level is a separable function of row, column, FGS-frame clock sequence number (frame-CSN) and FGS-parallel clock sequence number (parallel-CSN). The algorithm allows fitting any combinations of frame-CSN and parallel-CSN, but we have limited the parallel-CSN modeling range to pixels 1–29 and 565–66. The range 30–564 has not exhibited significant time-varying features. The frame-CSN pixels 18–20 are only measured in reversed-clocked cadence data, and are monitored but assumed constant.

The clock states included in DYNABLACK analysis are set by the following Dynablack Module Parameters (defaults shown in MATLAB syntax):

These parameters are used in the A1 fit.

- `parallelPixelSelect: [565, 566, 1:29]`
- `framePixelSelect: [1:16]`

These parameters are used in the A2 fit.

- `a2ParallelPixelSelect: [565, 566, 1:29]`
- `a2FramePixelSelect: [1:20]`

Regions excluded from the fits are those potentially affected by moiré pattern drift changes or undershoot produced by signals from bright stars near the trailing black. These appear to be caused by temperature changes induced by the signals from bright stars on the readout circuit. For the brightest stars, the signal may bleed over many rows and therefore the scene-dependent region extends

over many rows. The algorithm used for selection of scene dependent exclusion zones uses FFI images and depends on choices of: a) a signal threshold in DN read<sup>-1</sup>, above which a pixel may cause a scene-dependent artifact, b) a column threshold, above which a signal-threshold-crossing pixel may cause a scene-dependent artifact in the trailing black, and c) the row pad count to extend each contiguous region, which accounts for potential variations in star signal levels. If a row of a robust averaged (cosmic ray cleaned) FFI contains a pixel above the signal threshold at a column beyond the column threshold, then any row within the pad count of that row will be excluded from the trailing black fits (Thompson et al., 2016, Section 2.3.5.5).

The signal threshold, column threshold and row pad count are set in DYN-ABLACK by the following module parameters (defaults shown):

- scDPixThreshold: 5000
- nearTbMinpix: 1000
- blurPix: 1

Only specific regions of an image are suitable for determining spatial coefficients. To systematize the selection of these regions, the algorithm accepts specifications of a series of rectangular Regions of Interest (ROI), which are basically building blocks for the complete modeled image region, containing pixel or collateral point attributes needed to assemble the modeled response vectors and design matrices. These are summarized in Table 4.2.

**Table 4.2** Regions of Interest containing data used for black level spatial fitting. The “type” column simply identifies the source of response vector data, “ARP” means artifact removal pixels from target LC data, “Collateral” means summed collateral data, “FFI” means FFI pixel data, “RCLC” means target pixel data from reverse-clocked long cadences. The target table that specifies RCLC pixels was designed to monitor the FGS crosstalk, RBA and MPD artifacts (Van Cleve & Caldwell, 2016, Section 6.8).

Region of Interest (ROI)	min row	max row	min col.	max col.	type	main use
Leading ARP	7	1059	3	12	ARP	▪ row dependence
Trailing ARP	7	1051	1115	1132	ARP	▪ FGS crosstalk
Trailing Black Collateral	7	1059	1119	1132	Collateral	▪ undershoot
Trailing FFI	7	1063	1113	1132	FFI	
Masked Smear Collateral	7	18	13	1112	Collateral	▪ column dependence
Virtual Smear Collateral	1047	1058	13	1112	Collateral	▪ Start of Line Ringing (SOL)
Reverse-Clocked Long Cadence	7	1058	3	1130	RCLC	▪ FGS crosstalk

The complete model is assembled from a set of components. To systematize the construction of these components, we identify representative, distinct namable terms in the model that can be thought of as building blocks for the complete modeled design matrix and their associated indices for exclusion of potentially scene-dependent rows as a function of channel and all-zero columns in the design matrix. These are essentially the components described at the beginning

**Table 4.3** The following DYNABLACK module parameters control the Regions of Interest (defaults shown). All 4-element vectors are organized as [min row, max row, min column, max column]. All CCD row and column indices are 1-based:

Region of Interest	[min row	max row	min column	max column]
leadingArp:	[7	1059	3	12]
trailingArp:	[7	1051	1115	1132]
trailingArpUs:	[1052	1063	1113	1132]
trailingCollat:	[7	1059	1119	1132]
neartrailingArp:	[1057	1063	1100	1112]
trailingFfi:	[7	1063	1113	1132]
rclcTarg:	[7	1058	3	1130]
trailingMaskedSmear:	[1	20	1113	1132]
leadingMaskedSmear:	[1	20	1	12]
a2SolRange:	[1:280]			
a2SolStart:	13			

of this section. There are also “delta” components that measure the difference between leading and trailing black or masked and virtual smear as detailed in Appendix 4-A. The design matrix is constructed with one row of information for each element of acquired data within each ROI model. Prior to performing the cadence-by-cadence fitting for a given channel, we fit the pixel-by-pixel average of all specified cadences to obtain the exponential time constant parameter for the row-dependent exponential term by using a nonlinear model. Then we perform a linear fit on the mean of pixel values over selected long cadences to produce a set of mean coefficients. The resulting exponential time constant is used for all single-cadence fits for a given channel.

Finally, we perform cadence-by-cadence fits to obtain linear coefficients for individual cadences. The algorithm consists of the following steps.

For each channel and for each long cadence:

1. Select the response vector data from the input raw long cadences using the ROI indices described above.
2. Determine the cadence-specific undershoot component of the design matrix, which we apply to only a small subset of the modeled pixels, as described in Appendix 4-A.
3. Concatenate the constant part of the design matrix with the undershoot component.
4. Perform a linear fit of scene-dependent-free data to obtain linear coefficients, and store results in output structures.

For each channel and for each FFI:

1. Select the response vector data from the input raw FFI using the ROI indices described above.

2. Determine the FFI-specific undershoot component of the design matrix, including all modeled pixels.
3. Concatenate the constant part of design matrix with the undershoot component.
4. Perform linear fit of all data to obtain linear coefficients, and store results in output structures.

### 4.3.2 Thermo-Temporal Fitting

The spatial fit for each cadence produces a time series of spatial coefficients. These coefficients exhibit a variety of behaviors, which are often, but not always, simple functions of time or temperature. The purpose of the thermo-temporal fitting algorithm is to fit the time series of each spatial coefficient,  $C$ , to the equation:

$$C(r, T) = K_0 + K_t t + K_T T, \quad (4.2)$$

where  $K_0$  is a constant,  $K_t$  is the linear trend in  $C$ ,  $K_T$  is the temperature coefficient of  $C$ ,  $t$  is time, and  $T$  is temperature.

In addition to fitting the model described by C where  $K_0$ ,  $K_t$  and  $K_T$  are free parameters, the algorithm also fits the three constrained models where  $K_t = 0$ ,  $K_T = 0$  and  $K_t = K_T = 0$  and provides the necessary statistics to evaluate which models are consistent with the data based on the  $\chi^2$ . DYNABLACK also provides an option for adding step discontinuities to all four models at any data gaps longer than some cadence gap threshold.

The following DYNABLACK module parameters control whether or not to include step discontinuities in the thermo-temporal models (defaults shown):

- `includeStepsInModel`: true
- `cadenceGapThreshold`: 2

### 4.3.3 2-D Black Correction

Not all the information from the spatial and thermo-temporal fits is typically required to adequately correct the data for the observed artifact-induced variations. Some of the fit parameters are intended only for monitoring instrument performance. The 2-D black correction only includes the following terms.

$$X_{RC} - Z = X_0 + f_R + g_C + W_{RC}^F + W_{RC}^P, \quad (4.3)$$

where  $X_0$  is a constant,  $f_R$  is the row dependent term,  $g_C$  is the column dependent term,  $W_{RC}^F$  is the FGS frame crosstalk-dependent term, and  $W_{RC}^P$  is the FGS parallel crosstalk-dependent term:

The “undershoot” and the “delta” components are treated as static and are monitored. The equation is separable in terms of the row, column and crosstalk

dependent parts, which enables the 2-D corrections to be calculated from only the four vector terms, significantly shortening the processing time compared with an inseparable function. These terms separate as follows: vertical coefficients, horizontal coefficients, FGS-frame crosstalk coefficients and FGS-parallel crosstalk coefficients.

The retrieval algorithm evaluates the fitted model and returns the black value as a function of row, column and long cadence. For short cadence processing in CAL, fractional LC numbers are accepted and the blacks values returned are the linearly interpolated values between the adjacent integer long cadences. The model is assembled from the spatial and thermo-temporal fit coefficients based on a decision tree. If the  $\chi^2$  of the thermo-temporal fit indicates that the time series of the given spatial coefficient is not excluded from being a specific model at the 95% confidence level, then the spatial coefficient is determined from that model for all specified long cadences. In the event that multiple models meet the criterion, the preference order for the models, from highest to lowest is: 1)  $K_t = K_T = 0$ , 2)  $K_t = 0$ , 3)  $K_T = 0$  and 4)  $K_t$  and  $K_T$  are free parameters. If none of the cases meet this criterion, a comparison between the standard deviation of the differences between consecutive coefficients in the time series and the spatial coefficient fit errors is made. If the standard deviation of the differences is more than 1.5 times larger than the standard errors in the individual coefficients, as determined in the least-squares spatial fits, then the coefficients are applied discretely. Otherwise they are smoothed using a quadratic thermo-temporal fit over an adaptive time interval. Based on empirical data, the smoothing option is not available in DYNABLACK to the vertical coefficients, only to the horizontal, the FGS-frame crosstalk and FGS-parallel crosstalk coefficients.

The thermo-temporal may be determined automatically as described above, set to one of the four models or set to no model (raw or smoothed coefficients). These options are selected when retrieving the dynamic black fit in CAL.

The following CAL module parameters control the Dynamic Black thermo-temporal model selection (defaults shown):

- dynoblockModelAutoSelectEnable: true
- coefficientModelId: N/A
- dynoblockChi2Threshold: 0.9500

Note that a particular model may be selected by setting the following CAL module parameters:

- dynoblockModelAutoSelectEnable = false
- coefficientModelId = {-1, 1, 2, 3, 4}
- coefficientModelId = 1 →  $K_t = K_T = 0$
- coefficientModelId = 2 →  $K_t = 0$

- coefficientModelId = 3 →  $K_T = 0$
- coefficientModelId = 4 →  $K_t$  and  $K_T$  are free parameters
- coefficientModelId = -1 → raw or smoothed coefficients.

#### 4.3.4 RBA Flagging

The RBA and MPD share the same source signal, the high-frequency amplifier oscillation described above; however RBA are somewhat simpler to identify. The effects of RBA on pixel time series are generally larger and may be a greater risk to complicate the search for planet signatures. For these reasons, we have developed separate algorithms for detection and flagging of each. Only the RBA flagging algorithm has been implemented as part of DYNABLACK in the *Kepler* pipeline. The RBA flagging algorithm is summarized in Table 4.4.

**Measurement:** The signature of a rolling band is a time-varying displacement in trailing black spatial fit residuals, not due to scene dependent artifacts. The algorithm searches for these on a row-by-row basis for each channel. It also convolves a square wave transit kernel with these time series, as well as the column-by-column difference between masked and virtual smear time series for places where the black level variations exhibit transit-like signature. We call these “black transits.” Clearly, when the collateral data exhibits transit-like time signatures, the data in that row should be flagged, even if the typical variation is small. The convolution is equivalent to a least-squares fit to a transit signature centered at each long cadence. The fit transit depth is treated as a bias, and the uncertainty in the transit depth determines the noise level. The convolution result normalized by the noise level is the detection statistic.

**Detection:** The RBA flagging algorithm compares measured detection statistics with acceptability thresholds to detect unacceptable artifact behavior. As mentioned in section Subsection 4.2.3, the thresholds for flagging bias variations are based on limiting changes to  $0.02 \text{ DN read}^{-1} \text{ pixel}^{-1}$ . The result is a map determining which parts of row/cadence space exceed the threshold criteria. Detection statistics are filtered to clean spuriously flagged regions and a padding algorithm adds a buffer zone in row/cadence space around areas with a high density of above-threshold flags. The former prevents unnecessary flagging of data that is not at significant risk to interfere with planet search algorithms. The latter accounts for the fact that we do not directly measure the science pixels but rather collateral data that is indicative of the science pixels, and thus there is some level of uncertainty in boundaries of affected locations. The result is a Boolean map in row/cadence space indicating which regions are at risk.

**Severity evaluation:** The RBA flagging algorithm assembles a set of statistics that characterize the severity of the flagged artifacts in the identified regions. The detection statistic for the flagged regions is compared to the detection threshold and a 2-bit severity level is produced, indicating severity at the 1–2, 2–3, 3–4, and  $> 4 \times$  threshold level.

RBA detection is executed on several time scales by performing the convolution with square wave kernels of various integer long cadence lengths. The Boolean RBA flags at each time scale (test pulse duration) are packaged with both the severity parameters and their measured detection statistic to form a list. This is the primary output of the RBA flagging algorithm.

The following DYNABLACK module parameters control the RBA flagging (defaults shown):

- `cleaningScale`: 21
- `meanSigmaThreshold`: 1
- `pixelNoiseThresholdAduPerRead`: 1.65
- `pixelBiasThresholdAduPerRead`: 0.0160
- `robustWeightThreshold`: 0.5
- `severityQuantiles`: [0.9770 0.5000]
- `testPulseDurations`: [3 6 12 24 31]
- `transitDepthSigmaThreshold`: 0

## 4.4 Results

We applied these algorithms to a cross-section of *Kepler* flight data. The following paragraphs describe the results of these prototype runs. Table 4.5 identifies the flight data used to exercise each algorithm. The channels 2.1 and 12.1 were selected as “good channels”, 6.2 and 20.2 exhibit worst case FGS crosstalk, and 9.2 and 17.2 exhibit worst-case MPD.

### 4.4.1 Example Fits

*Kepler* readout channels exhibit a wide variety of black level morphologies. This section illustrates the model response fidelity using a few typical channels that span this range. Figure 4.5 shows the raw trailing black collateral data for two channels, 12.1 and 20.2 along with the fit curves. Some row regions were excluded from the fits because bright stars near the trailing black would have introduced un-modeled scene-dependent artifacts that could bias the fit coefficients if not explicitly removed.

The variation in the low frequency row-dependent terms of the model is evident from the examples in Figure 4.5. The series of spiked rows, repeating five times, are collateral rows containing FGS crosstalk sensitive pixels. The fit estimates offset values for each crosstalk-sensitive pixel from the combination of

**Table 4.4** Summary of RBA flagging algorithm. Numerical values specified are nominal values for algorithm parameters.

ITEM	RBA
Measurement	<ul style="list-style-type: none"> <li>▪ time-varying displacement in trailing black spatial fit residuals, not due to scene-dependent artifacts</li> <li>▪ black transit search in trailing black residuals and smear differences</li> </ul>
Detection	<ul style="list-style-type: none"> <li>▪ displacement sigma vs. time <math>&gt; 0.02 \text{ DN/pixel/read}</math></li> </ul> <p style="text-align: center;">—or—</p> <ul style="list-style-type: none"> <li>▪ both fixed residual displacement and black transit <math>&gt; 0.02 \text{ DN/pixel/read}</math></li> </ul>
Filtering	<ul style="list-style-type: none"> <li>▪ filter out flag densities less than 5% in 10 row by 39 LC region</li> <li>▪ pad around surviving flags <math>\pm 19 \text{ LC}</math> and <math>\pm 10 \text{ rows}</math></li> <li>▪ package flagged regions into rectangular suspect data flags (SDF)</li> </ul>
Severity Evaluation	<ul style="list-style-type: none"> <li>▪ Calculate statistics of measured parameters within each SDF region.</li> <li>▪ 97.7 percentile &amp; median displacement and fraction of exposure exceeding displacement threshold</li> <li>▪ 97.7 percentile &amp; median noise and fraction of exposure exceeding noise threshold</li> <li>▪ 97.7 percentile &amp; median bias variation and fraction of exposure exceeding bias variation threshold</li> <li>▪ number of <math>&gt;3</math> &amp; <math>&gt;4</math> sigma transit-like features</li> <li>▪ total fraction of exposure exceeding combined thresholds</li> </ul>

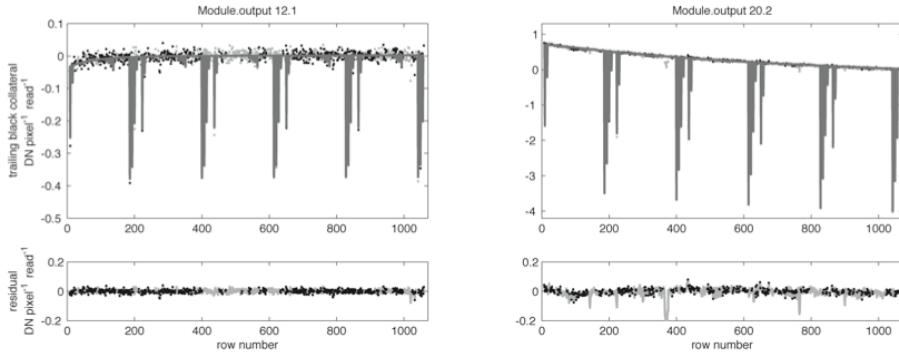
**Table 4.5** Module Outputs and time intervals used to exercise the various algorithms. Q0 lasted 10 days during commissioning from May 1–11, 2009; Q1 lasted 34 days from May 13–June 15, 2009.

Item	All Long Cadences Q0–Q1	All LC first 10 days of Q1	Decimated set: every 10th LC for Q0; every 20th LC for Q2
Fitting	2.1, 6.2, 9.2, 12.1, 17.2, 20.2	All	All
Correction	2.1, 6.2, 9.2, 12.1, 17.2, 20.2	None	None
RBA Flagging	2.1, 6.2, 9.2, 12.1, 17.2, 20.2	All	None
MPD Flagging	2.1, 6.2, 9.2, 12.1, 17.2, 20.2	All	None

both the collateral row values representing the sum of 14 pixels and the individual ARPs. Undershoot coefficients represent filter coefficients for an additive term to each pixel based on a linear combination of the values of the previous 20 (not shown). Figure 4.6 shows typical time dependent behavior of low-frequency serial and FGS-frame coefficients. FGS-parallel coefficients are similar.

#### 4.4.2 Effect of Corrections on Targets

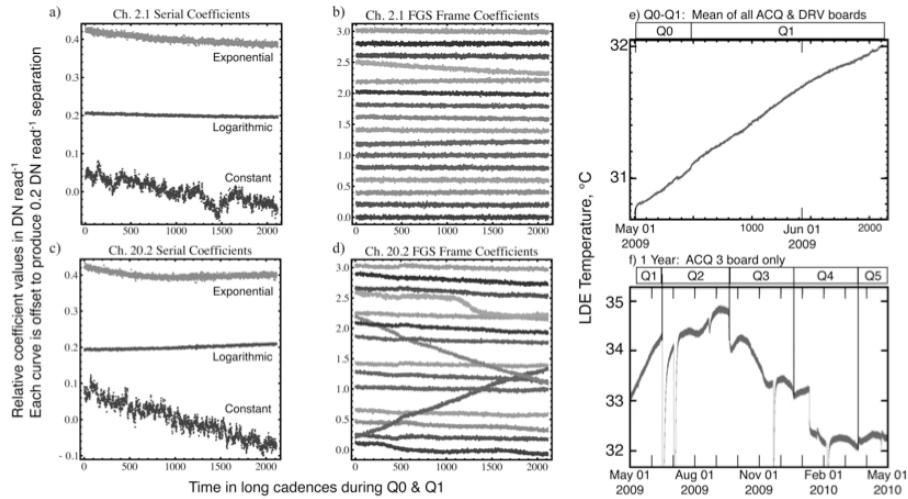
Calibrated *Kepler* pixels are subject to three separate effects of FGS crosstalk stemming from a) the 1-D black correction, b) the smear correction and c) the



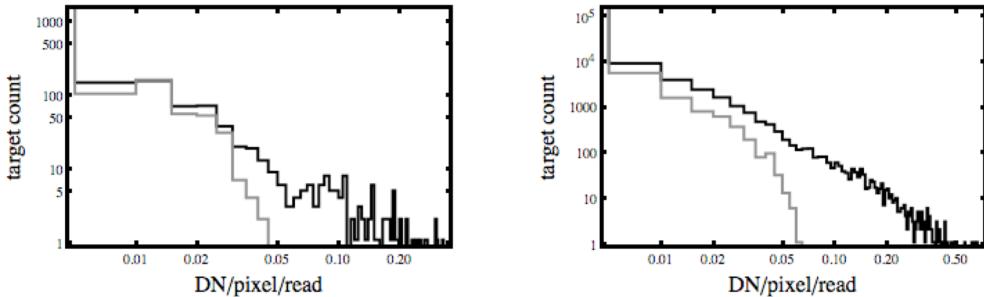
**Figure 4.5** Example comparison of fit curves (dark gray) and data (black) comparing trailing black collateral for channels 12.1 and 20.2 for one representative Q1 long cadence. The scales include an arbitrary constant offset. The row dependence of the black level is modeled well by an exponential term with a time constant that varies from channel to channel plus a logarithmic term which is channel independent. The light gray points are data and residuals from regions excluded from the fit because of increase likelihood of scene-dependent bias due to stars with pixel values  $>5000 \text{ DN read}^{-1}$  within 400 columns of the trailing black in the excluded rows. The density of stars is higher in channel 20.2 in Q1, so the likelihood of stars very close to the trailing black is higher, leading to the evident higher number of obvious outliers in the excluded region in that channel. From Figure 4 of Kolodziejczak et al. (2010).

effects of any FGS-cross-talking pixels in the target aperture. In the crosstalk-unaware 1-D black correction, a robust row-dependent fit is applied to the trailing black collateral data. Even with a robust fitting algorithm, there is likely to be some time-varying bias introduced by FGS crosstalking pixels with nonzero weights on the fit coefficients and thereby on the correction of each science pixel. A total of 89% of collateral rows are unaffected by FGS-cross-talking pixels. The introduced bias in a given row is applied uniformly to all science pixels in that row. The crosstalk-unaware smear correction is a cadence-by-cadence column-dependent correction based on measurements from the masked and virtual smear collateral regions. Each smear region is the sum of 12 rows. A total of 57% of collateral columns are unaffected by FGS-cross-talking pixels, and 23%, 3%, 12%, 5% have  $\{4, 3, 2, 1\}$  modeled pixels in a given column. The introduced bias in a given column is applied uniformly to all science pixels in that column. Both effects are reduced by the averaging and filtering afforded by the robust fits for the 1-D black correction and the averaging over typically 24 rows for the smear correction.

The FGS crosstalking pixels in the science pixel region represent the primary motivation for a dynamic 2-D black correction. Target apertures currently have no time-varying correction for crosstalking pixels. 82% of targets contain no FGS crosstalking pixels. In this case the introduced bias in a given pixel applies to only science pixels in that target aperture. The overall impact on the affected 18% of targets is substantially reduced by averaging over the whole aperture, and by the fact that the crosstalk is somewhat smaller and changes sign for some pixels, thereby producing further dilution by averaging.



**Figure 4.6** Examples of behavior of spatial fit coefficients vs. time for channels 2.1 (a,b) and 20.2 (c,d). a) and c) show the behavior of the low-frequency spatial coefficients. The bottom traces show the constant term, which is the only coefficient that always varies discretely (i.e., cannot be smoothed). The middle and top trace are the log and exponential coefficients, respectively. b) and d) show the FGS frame crosstalk coefficients. The scales are DN pixel<sup>-1</sup> read<sup>-1</sup> vs. time as measured by long cadence number. An arbitrary offset has been added to each curve to provide visual separation. LDE temperature variation is shown in e) for Q0–Q1 only and f) for an entire year. The LDE temperatures in e) are derived from the mean of temperatures measured on all five LDE board pairs (ACQ & DRV), whereas f) shows only the ACQ 3 board temperature. Temperature and time were strongly correlated during Q0–Q1, but less so for the other quarters. Spacecraft rolls at quarter boundaries, except Q0–Q1, produce step changes, and the failure of module 3 in Jan-2010 produced a large step in ACQ 3, which is used for module 3 readout. Safe modes explain the remainder of the discontinuous features in f). Temperature variations are generally  $\sim 1^{\circ}\text{C}$  within a quarter, and  $\sim 2^{\circ}\text{C}$  over the year, excluding the step due to module failure. From Figure 5 of Kolodziejczak et al. (2010).

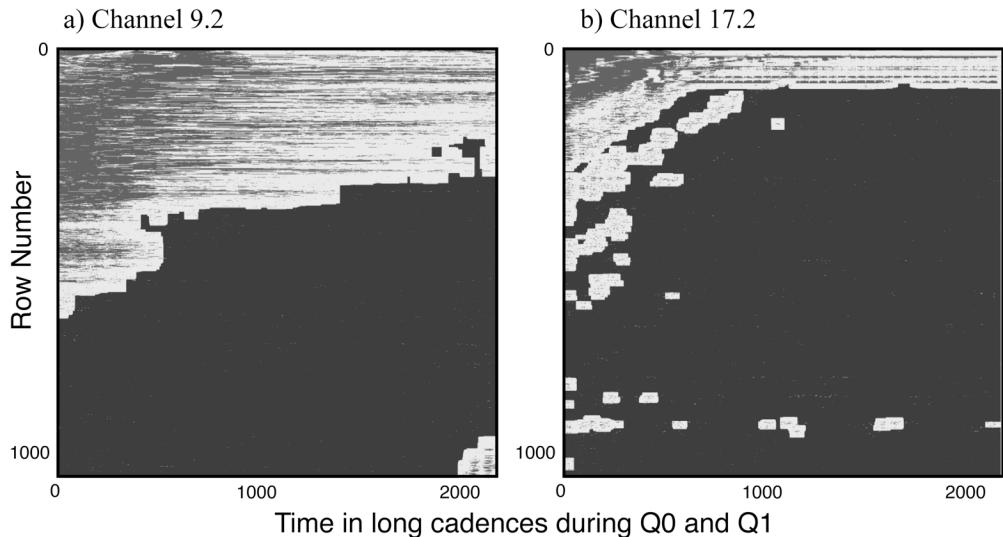


**Figure 4.7** Histograms of peak-to-peak dynamic 2-D black corrections per science target for a) worst case channel 20.2 and b) all channels during Q1. The contribution of collateral effects is shown in gray. From Figure 6 of Kolodziejczak et al. (2010).

Figure 4.7 shows a histogram of the peak-to-peak dynamic 2-D black corrections per target for Q1 where mean LDE temperature varied by  $1^{\circ}\text{C}$  over the full 34-day interval. The figure shows both a worst case channel, which is highly susceptible to crosstalk, and the cumulative impact on all channels. The gray histogram shows the separated effect of FGS-crosstalk in the collateral regions. The combined local pixels plus collateral (collateral only) affect 13% (5%) of

targets by more than  $>0.02 \text{ DN pixel}^{-1}\text{read}^{-1}$  in the worst case channel, while they affect only 4% ( $>1\%$ ) of all targets at this threshold level. The potential exists for a larger fraction of targets to be affected by  $>0.02 \text{ DN pixel}^{-1}\text{read}^{-1}$  for full quarters, but because the temperature variation is large during this period, these values are likely to be typical of a quarter. It is noteworthy that the worst case average temperature coefficient for a target is  $\sim 0.6 \text{ DN pixel}^{-1} \text{ }^{\circ}\text{C}^{-1}$  implying that a  $0.03 \text{ }^{\circ}\text{C}$  change in LDE temperature with the appropriate time signature would be required to produce the signature of an Earth-size planet. Even then, the thermal excursion would need to repeat three times at regular intervals to produce a false positive. A more likely possibility is that, left uncorrected, the less predictably varying pixels, as shown in Figure 4.6d, would reduce sensitivity to transits for limited periods in a small fraction of targets.

#### 4.4.3 Flagging Effectiveness



**Figure 4.8** Example of rolling band flagging results from channels a) 9.2 and b) 17.2 for combined Q0-Q1. Varying shaded areas represent the degree beyond threshold level that the data indicate a rolling band, white regions are the padding around the offending regions and black areas are unflagged. The scales are image row vs. time as measured in long cadence periods. From Figure 7 of Kolodziejczak et al. (2010).

Two examples are shown in Figure 4.8 of rolling bands detected and flagged by the algorithm described above for the result of combining Q0 and Q1. The varying shaded areas represent the degree beyond threshold level that the data indicate a rolling band, the white regions are the padding around the offending regions and the black areas are unflagged. The figure shows the movement of the rolling bands as the flagged regions change with time. The pattern in the signal over time at the top of Figure 4.8b is from a scene-dependent region caused by a variable star near the trailing black region. When a scene-dependent region is flanked by a rolling band, the flagging algorithm automatically flags the scene-

dependent region as part of the rolling band. If no rolling band is present, the flag remains scene dependent. The rolling band flags apply to all columns in a flagged row so we can define the rolling-band-free source coverage as the fraction of black cells in these 2-D maps. Here we define “artifact-free source coverage” as the fraction of the available field-of-view solid angle and exposure time that is unaffected by a given artifact. In the test cases below the rolling-band-free source coverage is 60% and 78% for module-outputs 9.2 and 17.2, respectively. The rolling-band-free source coverage was 100% for all but 10 channels in the first 10 days of Q1 data.

## Appendix A: Terms in the Spatial Model

This appendix defines the terms in the spatial model. We are modeling several discrete components with only the leading ( $C \leq 12$ ) and trailing black so it will be convenient to define the following discrete delta functions:

$$\delta_{XY} = \begin{cases} 1 & \text{if } X = Y, \\ 0 & \text{if } X \neq Y \end{cases} \quad (\text{A.1})$$

$$\delta_{C \in LB} = \begin{cases} 1 & \text{if } C \leq 12, \\ 0 & \text{if } C > 12 \end{cases} \quad (\text{A.2})$$

$$\delta_{R \in MS} = \begin{cases} 1 & \text{if } R \leq 20, \\ 0 & \text{if } R > 20 \end{cases} \quad (\text{A.3})$$

$$\delta_{R \in VS} = \begin{cases} 1 & \text{if } R \leq 1045, \\ 0 & \text{if } R > 1045 \end{cases} \quad (\text{A.4})$$

$$\delta_{\{R,C\} \in US} = \begin{cases} 1 & \text{if } \{R, C\} \text{ undershoot ARPs,} \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.5})$$

$$\delta_{\{R,C\} \in i}^F = \begin{cases} 1 & \text{if } \{R, C\} \text{ pixels with FGS frame sequence number } i, \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.6})$$

$$\delta_{\{R,C\} \in i}^P = \begin{cases} 1 & \text{if } \{R, C\} \text{ pixels with FGS parallel sequence number } i, \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.7})$$

With these we can define the various modeled terms for LC fits:

$$f_{RC} = \left( C_1^f + \delta_{C \in LB} C_2^f \right) \log \left( \frac{R}{R_{\log}} + 1 \right) + \left( C_3^f + \delta_{C \in LB} C_4^f \right) e^{-\frac{R}{R_{\exp}}}, \quad (\text{A.8})$$

where  $R_{\log}$  and  $R_{\exp}$  are constants;

$$g_C = \sum_{i \in C_{LB}} \delta_{Ci} C_i^g + \delta_{C \in LB} C_0^g, \quad (\text{A.9})$$

where  $C_{LB}$  is the list of discretely modeled leading black columns;

$$W_{RC}^F = \sum_{i \in FGS-F} \delta_{\{R,C\} \in i}^F \left( C_{1,i_{FGS-F}}^{WF} + \delta_{C \in LB} C_{2,i_{FGS-F}}^{WF} \right), \quad (\text{A.10})$$

where  $FGS - F$  are modeled FGS frame clock states;

$$W_{RC}^P = \sum_{i \in FGS-F} \delta_{\{R,C\} \in i}^P \left( C_{1,i_{FGS-F}}^{WP} + \delta_{C \in LB} C_{2,i_{FGS-P}}^{WP} \right), \quad (\text{A.11})$$

where  $FGS - P$  are modeled FGS parallel clock states; and

$$U_{RC} = \delta_{\{R,C\} \in US} \left( C_0^U + \sum_{j=C-N_{US}}^{C-1} C_{C-j}^U X_{Rj} \right), \quad (\text{A.12})$$

where  $X_{Rj}$  is a measured pixel signal value, and  $N_{US}$  is the number of undershoot columns.

RCLCs are used to measure the column dependence of the black level using similar terms:

$$f_{RC} = U_{RC} = 0. \quad (\text{A.13})$$

$$\begin{aligned} g_{RC} = & \sum_{i \in C_{LB}} \delta_{Ci} (C_{1i}^g + \delta_{R \in MS} C_{2i}^g + \delta_{r \in VS} C_{3i}^g) + \delta_{C \in LB} [(C_{10}^g + \delta_{R \in MS} C_{20}^g + \delta_{R \in VS} C_{30}^g) + \\ & (C_{11}^g + \delta_{R \in MS} C_{21}^g + \delta_{R \in VS} C_{31}^g) C + (C_{12}^g + \delta_{R \in MS} C_{22}^g + \delta_{R \in VS} C_{32}^g)], \end{aligned} \quad (\text{A.14})$$

where  $C_{LB}$  is the list of discretely modeled leading black columns.

$$W_{i \in FGS-F}^F \delta_{\{R,C\} \in i}^F \left( C_1^{WF}, i_{FGS-F} + \delta_{R \in MS} C_{2,i_{FGS-F}}^{WF} + \delta_{R \in VS} C_{2,i_{FGS-F}}^{WF} \right), \quad (\text{A.15})$$

where  $FGS - F$  are modeled FGS frame clock states.

$$W_{i \in FGS-F}^P \delta_{\{R,C\} \in i}^P \left( C_1^{WF}, i_{FGS-F} + \delta_{R \in MS} C_{2,i_{FGS-F}}^{WF} + \delta_{R \in VS} C_{2,i_{FGS-F}}^{WF} \right), \quad (\text{A.16})$$

where  $FGS - F$  are modeled FGS frame clock states.

## Bibliography

- Argabright, V. S., VanCleve, J. E., Bachtell, E. E., et al. 2008. “The Kepler Photometer Focal Plane Array,” in Proc. SPIE, Vol. 7010, in Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter, 70102L
- Caldwell, D. A., Kolodziejczak, J. J., Van Cleve, J. E., et al., 2010. “Instrument Performance in Kepler’s First Months,” ApJL, 713, L92
- Chandrasekaran, H., Jenkins, J. M., Li, J., et al. 2010. “Semi-Weekly Monitoring of the Performance and Attitude of Kepler Using a Sparse Set of Targets,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401B
- Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al., 2010. “Initial Characteristics of Kepler Long Cadence Data for Detecting Transiting Planets,” ApJL, 713, L120
- Kolodziejczak, J. J., Caldwell, D. A., Van Cleve, J. E., et al. 2010. “Flagging and Correction of Pattern Noise in the Kepler Focal Plane Array,” in Proc. SPIE, Vol. 7742, High Energy, Optical, and Infrared Detectors for Astronomy IV, 77421G
- Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010. “Pixel-Level Calibration in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401X
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)



## CHAPTER 5

---

# PIXEL LEVEL CALIBRATIONS

BRUCE D. CLARKE<sup>1</sup>, DOUGLAS A. CALDWELL<sup>1</sup>, ELISA V. QUINTANA<sup>1</sup>, HEMA CHANDRASEKARAN<sup>1</sup>, JOSEPH D. TWICKEN<sup>1</sup>, JON M. JENKINS<sup>2</sup>, MILES T. COTE<sup>2</sup>, SEAN D. McCauliff<sup>3</sup>, TODD C. KLAUS<sup>4</sup>, CHRISTOPHER ALLEN<sup>5</sup>, AND STEPHEN T. BRYSON<sup>2</sup>

<sup>1</sup>The SETI Institute/NASA Ames Research Center, Mountain View, CA 94035, <sup>2</sup>NASA Ames Research Center, Moffett Field, CA 94035 <sup>3</sup>Wyle Labs/NASA Ames Research Center, Moffett Field, CA 94035  
<sup>4</sup>Stinger Ghaffarian Technologies, Inc./NASA Ames Research Center, Moffett Field, CA 94035, <sup>5</sup>Orbital Sciences Corporation/NASA Ames Research Center, Moffett Field, CA 94035

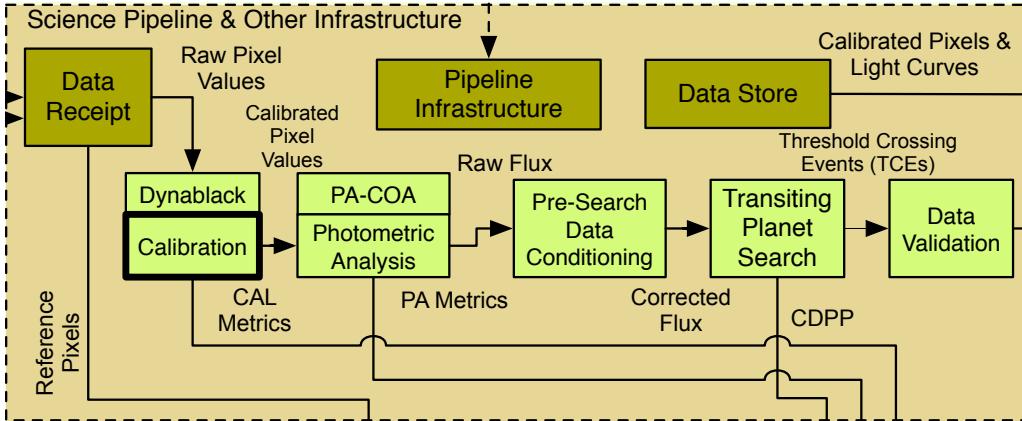
**Abstract.** This chapter describes the CAL science algorithms that calibrate pixel data by performing a series of corrections to the raw pixel data. Many of the primary corrections use external models particular to each CCD. These models were developed from pre-flight hardware tests and Full Frame Image (FFI) data collected during instrument commissioning (Haas et al., 2010) both in thermal-vacuum testing and on orbit prior to the dust cover ejection. We discuss how these models are applied to correct for the 2-D bias structure, gain and nonlinearity in the conversion from analog-to-digital units (ADU) to photoelectrons, local detector electronic effects (undershoot and overshoot) and spatial variations in pixel sensitivity (flat field). Instrumental effects that are compensated for in CAL include bleeding charge (excess charge from saturated stars that leaks along the CCD columns and potentially into the masked and virtual smear regions), outliers due to cosmic ray hits, thermally-activated dark current, spatially varying, thermally dependent bias voltage (or black) levels, and smearing of the image resulting from shutter-less readout of the CCD. This chapter is an significantly updated and revised version of Quintana et al. (2010).

**Keywords:** *Kepler* Mission, Transit Photometry, Calibration

### 5.1 Introduction

The task of calibrating the science data downlinked from the *Kepler* spacecraft is a challenging one, both in light of the electronic image artifacts discussed in Chapter 4 and because of the fact that only  $\sim$ 6% of the pixels from the focal plane are available for this purpose, due to the limited storage capacity of the Solid State Recorder (SSR) and the limited downlink bandwidth. In SOC 9.3, the Dynablack module analyses the trailing black measurements to formulate improved 1-D black corrections in the face of thermally dependent fine guidance sensor electronic crosstalk, and to identify regions of the CCDs and cadences affected by rolling band anomalies. The Calibration (CAL) module's task is to use the information provided by Dynablack as well as a suite of focal plane models

to calibrate on-chip artifacts, including those commonly encountered in CCD astronomy (e.g., flat field correction), and some artifacts peculiar to *Kepler* such as the readout smear incurred by the absence of a shutter and a finite readout time. Figure 5.1 shows where CAL fits in the context of the science data processing pipeline. The calibrated pixels furnished by CAL are used to extract brightness and location measurements from the target stars as the next steps in preparing the data for the transit search.



**Figure 5.1** CAL in the context of the architecture of the SOC. CAL performs the pixel level calibrations on the original *Kepler* science data using focal plane models and information from the Dynablack (DYN) module. The calibrated pixels are then processed by the Photometric Analysis (PA) module prior to conditioning the data for the transit search in Presearch Data Conditioning (PDC) to remove instrumental signatures and residual outliers.

In Section 5.2, we present an overview of the focal plane CCD array and the various pixel types that are calibrated. Section 5.3 describes the individual calibration steps, presented in the order that they are performed, along with the additional functionalities of CAL. In Section 5.4 we present a summary.

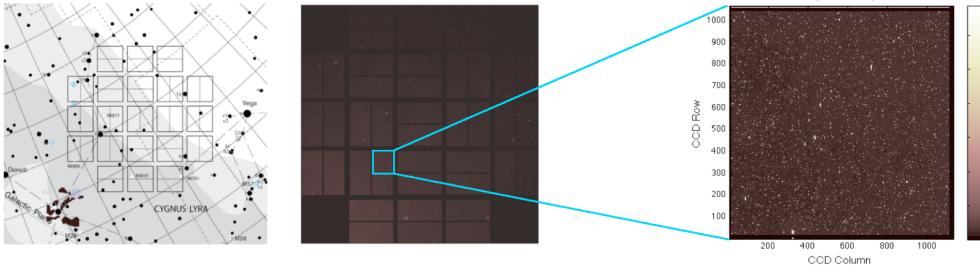
## 5.2 *Kepler* Data Formats and CAL Unit of Work

### 5.2.1 CAL Data Types: Long and Short Cadence and Full Frame Images

CAL is designed to operate on three data types that differ in the number of integrations that compose each sampling time (cadence) and in the number and location of the pixel data. The raw data include photometric pixels (target and background), along with a subset of pixels termed “collateral data” that includes masked and virtual (over-clocked) pixels around the perimeter of each CCD. The three types of data sets processed within CAL are: (1) Long Cadence Data (LC), select pixels from up to 170,000 long cadence targets collected every 29.4 minutes (with 270 exposures per cadence); (2) Short Cadence Data (SC), a subset of the LC pixels from 512 targets that are sampled more frequently, at 0.98 minute intervals (with 9 exposures per cadence); and (3) FFI data that contain all available pixels for a single long cadence.

### 5.2.2 Focal Plane Array

The *Kepler* focal plane array is composed of 42 CCD detectors (Figure 5.2; see Van Cleve & Caldwell (2016) for focal plane details). A CCD “Module” refers to a pair of CCDs that share a field flattener lens and are read out simultaneously by the detector electronics. Each of the 21 modules is composed of four CCD “Outputs” that are each read out by a separate analog signal chain.

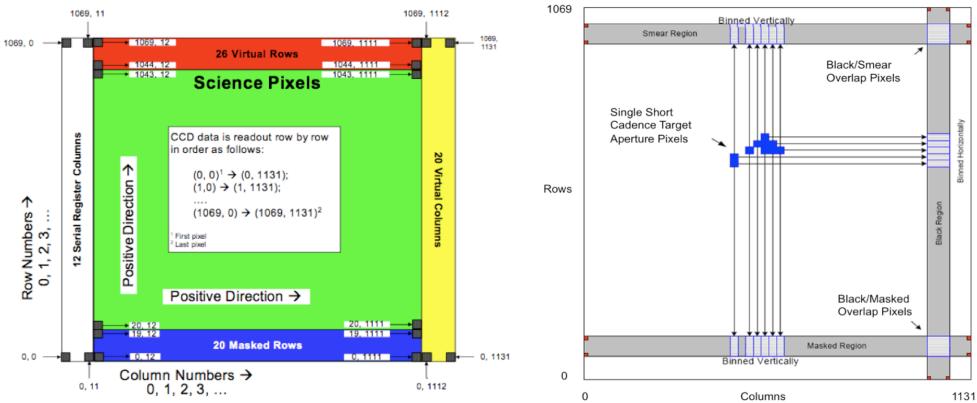


**Figure 5.2** A celestial view of the *Kepler* focal plane (left), the first light full frame calibrated image (middle), and a close-up of module/output = 17/2, channel 58 (right) are shown. The color axis on the images is in units of  $10^6$  photoelectrons. From Figure 2 of Quintana et al. (2010).

The science module labels are integers [2–4, 6–20, 22–24]. Note that the four fine guidance sensor modules on the corners of the focal plane [1, 5, 21, 25] are not listed as science modules. The output integers range from 1–4. There are 84 combinations of module/output numbers, and each is also referred to as a CCD channel and is specified by an integer from 1–84. Each channel can therefore be mapped to a unique module/output (e.g., channel 19 = module/output 7/3). The CAL software component operates on a single CCD module/output (channel) at a time.

### 5.2.3 Pixel Collection

Each CCD channel consists of an array of pixels with 1070 rows and 1132 columns, of which a  $1024 \times 1100$  subset is photometric pixels (Figure 5.3). For FFI data, the full  $1070 \times 1132$  array is downlinked whereas for LC and SC data, only select target and background pixels are downlinked due to limitations in onboard storage, communications bandwidth and flight software design (Bryson et al., 2010b). For LC data, an upper limit of 170,000 stellar targets and 94,500 background targets are collected across the focal plane with additional limitations on the number of targets per channel and the total pixel count. LC collateral data (black, masked smear and virtual smear pixels as described in the next section) are also collected for calibration. For SC data, a subset of the LC stellar targets is collected at a higher rate. There is a maximum of 512 SC targets across the focal plane. SC collateral data are also downlinked. The SC collateral data consist of the black pixels and smear pixels that lie in the projections of the SC target pixels onto the collateral regions plus the black pixels that lie in the projections of the masked and virtual smear pixels (see Figure 5.3).



**Figure 5.3** A schematic of the pixel regions in a single CCD channel (left) shows the location of photometric pixels, along with the collateral pixels on the perimeter of the CCD that are collected for calibration. Only a subset of collateral data (black columns and smear rows) is collected and co-added onboard the spacecraft. For LC data, all black rows (and a subset of columns) and all smear columns (and a subset of rows) are collected (gray region in right panel), whereas for SC only collateral data in the projections of the SC target pixels are collected. From Figure 3 of Quintana et al. (2010).

### 5.2.4 Photometric and Collateral Data

For both photometric and collateral pixels, the pixel values, row/column indices and logical data gap indicators for all cadences are presented to CAL. Photometric pixels are those in the unmasked region of the CCD and are referred to in CAL as the target and background pixels. Note that CAL does not distinguish between target pixels and background pixels, so all photometric pixels are calibrated in the same manner.

On each channel, the collateral data include the 12 leading black columns (virtual pixels read out before the photometric pixels in each row), 20 trailing black columns (virtual pixels read out after the photometric pixels in each row), 20 masked smear rows (the physical pixels closest to the serial register, which are covered with an opaque aluminum mask) and 26 virtual smear rows (virtual pixels read out after all of the photometric rows are clocked out). The down-linked collateral data include only co-added subsets of the collateral pixels for use in calibration. Downlinked black pixels for each CCD row are the result of co-adding 14 trailing black pixels in columns 1118–1131 (zero-based indices). The leading black pixels are not co-added or downlinked as collateral data due to the presence of image artifacts in that region; however, individual pixels in the leading black are downlinked as Artifact Removal Pixels (ARP) targets and are used in the dynamic black estimates developed by the CAL sub-module Dynablack (see Chapter 4). Downlinked masked smear pixels for each physical CCD column (12–1111) result from co-adding the 12 pixels in masked smear rows 6–17. Downlinked virtual smear pixels for each physical CCD column result from co-adding the 12 pixels in virtual smear rows 1046–1057. There are 1100 co-added masked and virtual smear pixels corresponding to the physical columns (12–1111) per column per cadence.

For SC data, only a small number of targets are collected from each channel, and the rows and columns of each target determine which collateral pixels are collected (see Figure 5.3). Two additional pixel types are collected for SC: masked black pixels (the co-added sum of the pixels in the cross-section of co-added trailing black columns and co-added masked smear rows) and virtual black pixels (the co-added sum of the pixels in the cross-section of co-added trailing black columns and co-added virtual smear rows). Each masked black or virtual black pixel is therefore the sum of pixels in 14 black columns times 12 smear rows, or 168 co-adds per cadence.

For FFI data, the full  $1070 \times 1132$  pixel array for each detector channel is downlinked. CAL uses the spacecraft configuration map to determine which pixels from the collateral regions should be binned for calibration, and the data are then processed as if they were a single cadence of LC data.

Up to two small subsets of photometric pixels may be specially identified by row and column in CAL LC processing in order to track the 2-D black and LDE undershoot metrics. These subsets are called 2-D Black Targets and LDE Undershoot Targets.

### 5.2.5 Processing Order

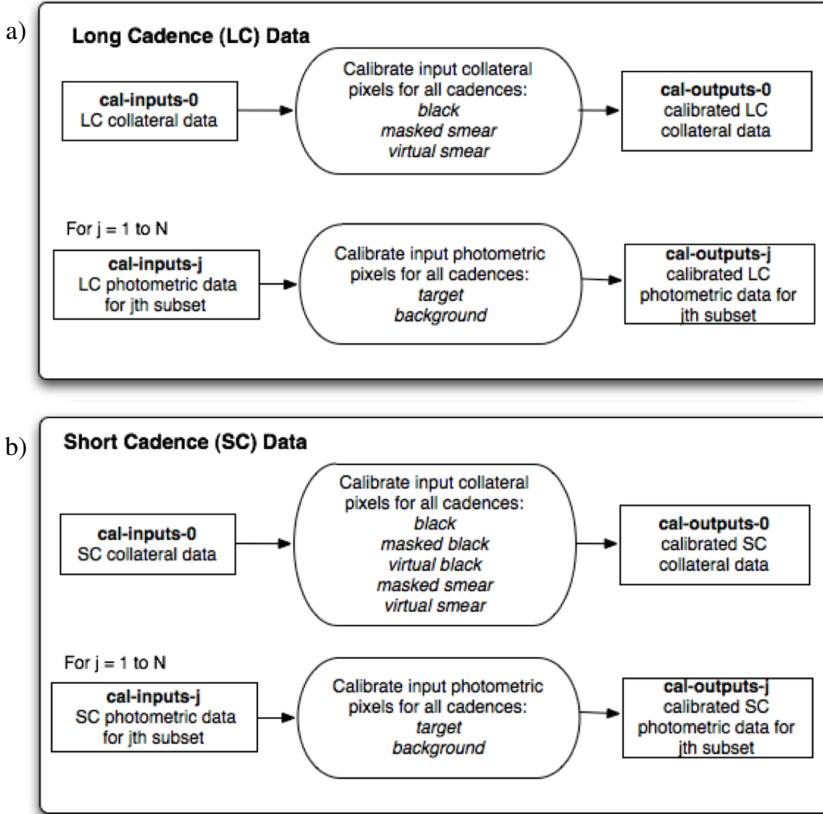
Data for each CCD channel are calibrated individually. Regardless of data type (LC, SC, or FFI), the collateral pixels are always processed first in order to estimate the black correction, smear correction and dark correction, respectively. The photometric pixels are calibrated next in multiple invocations. Due to the large volume of data, the target and background pixels are broken up into chunks, which include all the available pixels in a subset of the CCD rows. Each chunk is calibrated separately.

Figure 5.4 shows a high-level overview of the main CAL functions for SC data and LC data. All science algorithms are written in MATLAB (MAtrix LABoratory) software, and the main function that is called for each CAL invocation is referred to as the CAL MATLAB controller (`cal_matlab_controller.m`). This function receives all inputs via the Java pipeline infrastructure, calls the appropriate algorithms according to the data type (LC, SC, or FFI), and outputs the calibrated pixels and other CAL products back to the Java pipeline infrastructure. See Klaus et al. (2010) for complete details on this Java-MATLAB interface.

For each invocation, the `cal-inputs-*.mat` file contains a structure called *inputsStruct* that holds all data, and the following MATLAB command is executed for each invocation:

$$\text{outputsStruct} = \text{cal\_matlab\_controller}(\text{inputsStruct}) . \quad (5.1)$$

The *inputs* structure has the same fields for all CAL invocations, but the pixel data fields that are available (non-empty) determine which second-level functions to call from the controller (either the collateral or photometric main calibration functions). A cadence type flag is set in the *inputs* (for LC, SC, and FFI)



**Figure 5.4** Overview of the CAL MATLAB Controller for a) LC data, and b) SC data. The CAL MATLAB controller is first called using collateral data inputs, and then multiple invocations of the controller are called to calibrate the remaining photometric pixels, which are broken up into  $N$  chunks of data. The collateral data are packaged differently for LC and SC data.

in order to determine which functions (or parts of code within a function) are required to process each type. Apart from initial repackaging within CAL, FFIs are calibrated as a single cadence of LC data.

Table 5.1 and Table 5.2 further describe how the pixels are subdivided for the multiple invocations for each module/output and data type (LC, SC, and FFI). Note that filenames that end in “m” are MATLAB functions and those that end in “.mat” are MATLAB binary files that contain data.

The number of invocations is a pipeline parameter and the data type determines how the photometric pixels are subdivided. For LC data, each photometric pixel subset must include entire rows of pixels in order to properly correct for the LDE undershoot/overshoot, which involves filtering pixel flux along full rows. Each invocation includes all cadences available for the selected pixels. SC data sets include far fewer target/background pixels per CCD, but they are sampled 30 times more frequently than LC data. All cadences of all SC collateral pixels can therefore be processed in the first invocation, but the photometric pixels must be subdivided into chunks of cadences ( $\sim 5700$  cadences per chunk

**Table 5.1** Short Cadence CAL invocations, where the number of cadences,  $k$ , is a parameter that can be varied.

SC CAL			
Invocation	Inputs file	Inputs file contents	Outputs file
1	cal-inputs-0.mat	Collateral data (including masked and virtual pixels used to estimate the bias level, smear level, and dark current). Rows = all available Columns = all available Cadences = all available	cal-outputs-0.mat
2	cal-inputs-1.mat	A subset of the photometric pixels: Rows = all available Columns = all available Cadences = $\{1, \dots, k\}$	cal-outputs-1.mat
:	:	:	:
N	cal-inputs-(N-1).mat	A subset of the photometric pixels: Rows = all available Columns = all available Cadences = all remaining	cal-outputs-(N-1).mat

is nominal), which are processed in separate invocations. Note for FFIs, the full 2-D collateral regions (black, masked, and virtual smear) are inputs rather than just the binned 1-D arrays. In the FFI collateral invocation, the 2-D arrays are binned to look like LC data and then calibrated. The second invocation includes all pixels from the entire CCD array (collateral again plus photometric) which are then calibrated using the black, smear, and dark estimates from the collateral invocation.

### 5.2.6 Data Gaps

All pixel types (black, smear, and photometric) are accompanied by logical arrays the same size as the pixel arrays and indicate spatial and temporal gaps. Calibration steps are applied only to the available (non-gapped) pixels. The inputs to CAL include a `cadenceTimes` structure that includes timestamps, timestamp gap indicators, and spacecraft data quality flags (each is a  $\#$  cadences  $\times$  1 array). CAL uses timestamps that correspond to the middle of each cadence.

The spacecraft data quality flags used by CAL to augment the timestamp gaps are described below. Gapping on these flags may be enabled separately; however, the nominal configuration has been to run with all data quality flags enabled. The statements in parenthesis indicate the condition required to set a cadence gap, followed by the corresponding data archive QUALITY flag bit that is set in the light curve and target pixel FITS files (Thompson et al. (2016), Table 2–3).

**Table 5.2** Long Cadence CAL Invocations, where the number of rows,  $j$ , is a parameter that can be varied.

Long Cadence CAL Invocation	Inputs file	Inputs file contents	Outputs file
1	cal-inputs-0.mat	Collateral data (including masked and virtual pixels used to estimate the bias level, smear level, and dark current). Rows = all available Columns = all available Cadences = all available	cal-outputs-0.mat
2	cal-inputs-1.mat	A subset of the photometric pixels: Rows = $\{1, \dots, j\}$ Columns = all available Cadences = all available	cal-outputs-1.mat
:	:	:	:
N	cal-inputs-(N-1).mat	A subset of the photometric pixels: Rows = all remaining Columns = all available Cadences = all available	cal-outputs-(N-1).mat
N + 1	cal-inputs-(N).mat	No pixel input - The last invocation is used to aggregate performance metrics and diagnostics over invocations 1 through N.	cal-outputs-(N).mat

- **isMmntmDmp:** a momentum dump occurred during accumulation (isMmntmDmp = T, QUALITY bit #6)
- **isFinePnt:** spacecraft is not in fine point (isFinePnt = F, QUALITY bit #16)
- **isSefiAcc:** single event functional interrupt in accumulation memory (isSefiAcc = T, QUALITY bit #15)
- **isSefiCad:** single event functional interrupt in cadence memory (isSefiCad = T, QUALITY bit #15)
- **isLdeOos:** Local Detector Electronics out of sync reported (isLdeOos = T, QUALITY bit #15)
- **isLdeParEr:** Local Detector Electronics parity error occurred (isLdeParEr = T, QUALITY bit #15)
- **isScrcErr:** SDRAM Controller memory pixel error occurred (isScrcErr = T, QUALITY bit #15)

In addition to the spacecraft data quality flags there are exclude flags (exclude-Indicators) that are set by the pipeline operator at run time. Use of these

flags in CAL is determined by the CAL module parameters `enableExcludeIndicators` and `enableExcludePreserve`, which selects these gaps for the cosmic ray detector/corrector only. The nominal configuration is to use them only in CR detection and correction (`enableExcludePreserve = true`). The logical ‘or’ of all enabled spacecraft data quality flags relative to their gap condition, `excludeIndicators` (if enabled) and the original gaps in `cadenceTimes`, determines the cadence gap flags.

Gaps are further augmented in masked or virtual smear pixels that contain flux from saturated stars spilling along the columns into the collateral region. This bleeding charge would potentially contaminate the smear correction for all pixels in that column leading to poor calibration of the photometric pixels. Due to the careful positioning of the *Kepler* field of view to avoid extremely bright stars in the photometric pixel region, bleed into both the masked and the virtual smear regions for any particular column has been avoided for the most part, leaving the uncontaminated collateral region available for smear correction. However, to accommodate those cases where a bright saturated star is bleeding into either the virtual or masked smear, two bleeding column map files have been developed using a combination of FFIs and raw LC data. These are used to identify masked and virtual smear pixel gaps prior to calibration. Since the targets falling on a particular channel are season dependent there are a total of four mappings for each detector channel for both masked and virtual smear. These maps are stored as MATLAB m-files:

- `get_masked_smear_columns_to_exclude.m`
- `get_virtual_smear_columns_to_exclude.m`.

Note the column indices in these files are in 1-based units. The first CCD column begins with “1” and not “0”. The 1100 smear column indices are in the range 13–1112 (see Figure 5.3 and Thompson et al. (2016) and Sections 2.3.5.5, 2.3.5.6).

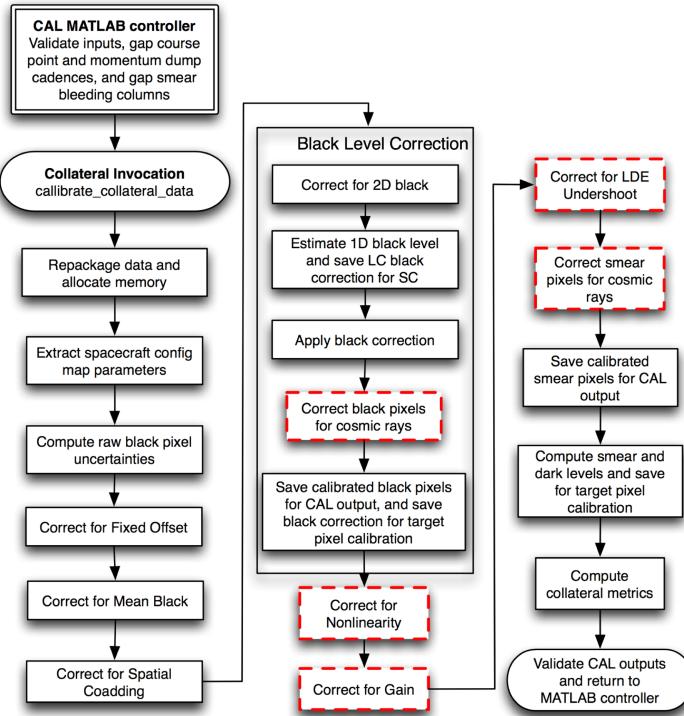
Pixels from gapped cadences are not calibrated. Some of the intermediate products (e.g., black and dark level estimation) may contain nearest neighbor interpolated values for cadences in which the gap indicators are set.

### 5.3 Calibration

The primary calibration steps are described in this section as well as Quintana et al. (2010). A schematic of the data flow in the CAL module is shown in Figure 5.5 and Figure 5.6. The boxes with dashed lines show the steps that can be disabled in the pipeline if desired. All cadence types (LC, SC, and FFI) and pixel types (collateral and photometric) are processed separately but use primarily the same MATLAB code base. Exceptions are noted in this section. There are a number of steps performed in the controller before and after the pixel level calibration is performed. The inputs are first validated to ensure that the available

data are within the appropriate bounds. All row and column fields are converted from java zero-based to MATLAB one-based indexing. If the Propagation of Uncertainties (`pouEnabled`) flag is set, a structure is created to store the primitive data and the kernel of each transformation performed on the data in turn. This structure is used to propagate uncertainties on the primitive data and produce calibrated uncertainties for the available pixel types.

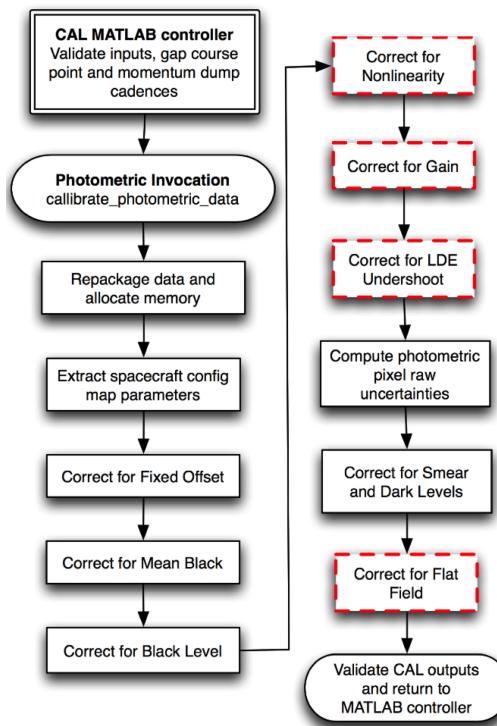
For each cadence type and channel, the first invocation calls the script `calibrate_collateral_data.m`, which processes collateral (black and smear) pixels for all cadences. The outputs from this first invocation include calibrated black and smear pixels, collateral and cosmic ray metrics, and the black, smear and dark current corrections that are needed to calibrate the photometric pixels. All remaining invocations call `calibrate_photometric_data.m` which calibrates the photometric pixels using output from the collateral invocation.



**Figure 5.5** Collateral data invocation. The steps with the red dashed borders can be enabled and disabled independently. However, they were enabled for all nominal *Kepler* science data processing.

The first few steps shown in Figure 5.5 and Figure 5.6 involve repackaging the input data, allocating memory, and retrieving parameters from the spacecraft configuration map. In the function `repackage_data_for_calibration`, data and parameters are collected and repackaged in formats that allow the code to be vectorized in order to improve run time performance. Flags are set here that indicate which type of data has been passed into CAL (`processLC`, `processSC`, `processFFI`), and which pixel types are available in the inputs (collateral or

target/background). All available pixel values, gaps, rows, and columns are saved in  $n_{pixels} \times n_{cadences}$  arrays. Short cadence data are saved as sparse arrays. Missing cadences are recorded for each type. These arrays are saved to *calIntermediateStruct*, which is passed through all subsequent functions in CAL. After each sequential correction, the pixel values and gap arrays in the intermediate structure are updated. Due to memory limitations, we do not save all arrays before/after each correction, but select intermediate arrays can be saved if a high debug level is set when running CAL on a development workstation for example. The debug level is always set to zero when running in the pipeline.



**Figure 5.6** Photometer data invocation. The steps with the red dashed borders can be enabled and disabled independently. However, they were enabled for all nominal *Kepler* science data processing.

The next step involves allocating memory for intermediate data produced within CAL (such as data used to compute uncertainties or metrics, or intermediate calibrated data saved with a high debug level), and preparing some of the output structures for the final calibrated data. These fields are also added to the *calIntermediateStruct* and are available to all subsequent functions.

Several additional parameters used in CAL are retrieved from the spacecraft configuration map (and added to *calIntermediateStruct*) in the script `get_config_map_parameters`:

- requantization table fixed offsets (see Van Cleve & Caldwell (2016)) §7.4.1)

- start/end columns of the black pixels that are co-added onboard the spacecraft (used to normalize *blackPixels* values)
- start/end rows of the masked and virtual smear pixels that are co-added onboard the spacecraft (used to normalize *maskedSmearPixels* and *virtualSmearPixels* values; both black columns and smear rows are used in SC data to normalize the *maskedBlackPixels* and the *virtualBlackPixels*)
- CCD read and exposure time
- number of exposures per cadence.

### 5.3.1 Compute Raw Black Uncertainties

The “black level” refers to the bias in each CCD module/output, an electronic offset that has been added to the CCD readout voltage to ensure that positive signals are input into the Analog to Digital Converter (ADC). The offset is  $\sim 5\%$  of the ADC input voltage range. Note that ‘black’ refers to this bias and should not be confused with the dark level, which is a measure of the CCD dark current. The black level has an intrinsic noise referred to as “read noise” or “readout noise”. In order to propagate the uncertainties from the primitive data, the uncertainties in the original black data (and masked/virtual black data for SC) are estimated at the beginning of CAL from the read noise model and quantization step size. First, the read noise in ADU is extracted from the focal plane characterization (FC) model. The quantization noise from the 14-bit ADC must also be taken into account and is equal to 1/12 (the variance of a unit-wide uniform random process):

$$\sigma_{\text{effectiveCoaddedReadNoise}}^2 = \sigma_{\text{readNoise}}^2 + 1/12. \quad (5.2)$$

The effective read noise calculated above is per pixel per read so it is scaled up by both the number of exposures per cadence and the number of pixels co-added to produce the *effectiveCoaddedReadNoise*.

The noise contribution due to requantizing the co-added pixel values prior to downlink is computed by extracting the quantization step size from the requantization table for the raw black pixels and using that step size to scale the unit-wide uniform process variance. The uncertainty is taken to be the square root of the sum of the variances:

$$\begin{aligned} \sigma_Q^2 &= \Delta_Q^2 / 12 \\ \sigma_{\text{raw black}}^2 &= \sqrt{\sigma_{\text{effectiveReadNoise}}^2 + \sigma_Q^2}, \end{aligned} \quad (5.3)$$

where  $\sigma_Q^2$  is the quantization noise due to the requantization step size  $\Delta_Q$  in ADU. These uncertainties are then renormalized by the number of co-added pixels so that their values represent the uncertainty per black pixel per cadence.

### 5.3.2 Models

Some of the calibration steps rely on external models used to characterize each CCD. These FC models (Allen et al., 2010) were developed from extensive ground-based testing, were updated in flight while the spacecraft dust cover was still in place, and are monitored in flight by pipeline metrics collected by the PPA (Photometric Performance Assessment) pipeline module. The FC models are available at the *Kepler* MAST archive and are described in Thompson et al. (2016), §2.3.5. The six time-dependent models used in CAL include:

1. A read noise model that gives the read noise per channel,
2. A 2-D black model that provides a 2-D map of the black/bias structure per channel,
3. A gain model that gives the ADU-to-photoelectrons conversion factor per channel,
4. A linearity model that provides a set of polynomial coefficients used in the gain transfer function to correct for nonlinearity in the readout amplifier;
5. An undershoot model that includes coefficients for the digital filter used to correct for undershoot/overshoot artifacts introduced by the CCD local detector electronics (LDE), and
6. A flat field model consisting of a 2-D map of values that are used to correct for pixel-to-pixel sensitivity.

Note that these FC models are allowed to be piecewise linear functions in time. If there are multiple entries in any model, the result of the correction should be the linear interpolation of the bracketing corrections for any pixel, where the time intervals between the time tag of the data and those of the prior and subsequent models determine the interpolation weights. That is, if  $t_1$  and  $t_2$  are the model times,  $t$  is the time of the data collection, and the two corrections applied to the data result in  $x_1$  and  $x_2$ , the corrected value is:

$$x = x_1 + \frac{(t - t_1)}{(t_2 - t_1)} (x_2 - x_1). \quad (5.4)$$

In addition to these models, which are maintained outside of CAL, there is also a set of global constants used throughout each pipeline module contained in the structure *fcConstants* that is included in *inputsStruct*.

The following flags are included in the CAL *inputsStruct* to enable or disable the noted corrections. These are fields of the *moduleParametersStruct*.

- *linearityCorrectionEnabled* – enable/disable the nonlinearity correction
- *undershootEnabled* – enable/disable the undershoot correction

- `crCorrectionEnabled` – enable/disable the cosmic ray correction
- `flatFieldCorrectionEnabled` – enable/disable the flat field correction
- `debugEnabled` – enable/disable the debug level
- `pouEnabled` – enable/disable the propagation of uncertainties (Note: this flag is a field of the `pouModuleParametersStruct`, also attached to the `inputsStruct`).

Additional CAL inputs worth noting are:

- `spacecraftConfigMap` – structure that contains mnemonics and values of configurable parameters that are uploaded to the spacecraft (see Subsection 5.2.1 for parameters needed by CAL; there may be multiple maps, each one applicable within specified timestamp intervals)
- `requantTables` – the requantization tables used to compute raw pixel uncertainties and to compute the theoretical and achieved compression efficiency (there may be multiple tables, each one applicable within specified timestamp intervals)
- `huffmanTables` – the Huffman tables used to compute the theoretical and achieved compression efficiency (there may be multiple tables, each one applicable within specified timestamp intervals).

### 5.3.3 Fixed Offset, Mean Black, and Spatial Co-Adds

The first step in calibration is to adjust the raw downlinked pixels for the fixed offset and mean black values that were applied prior to downlink. This adjustment is to compensate for variations in bias across the focal plane array and to accommodate the variations in temporal and spatial co-adding applied to the various data types. The black and smear measurements are spatially co-added and thus possess more read noise than the target and background pixels. The SC data have fewer temporal co-adds, and so have less read noise than the LC data. The fixed offset and mean black levels are used by flight software to shift the digital values to the same “zero point” for LC data and to a separate “zero point” for SC data (Jenkins & Dunnuck, 2011).<sup>1</sup> All pixels are subject to requantization in which each pixel value is mapped to a discrete value in a pre-generated table in order to control the A/D quantization noise (the round-off error resulting from digitizing the voltage signals) to within 1/4 of the intrinsic measurement uncertainty (Jenkins & Dunnuck, 2011). The `meanBlack` and `fixedOffset` adjust pixels on all channels to a common zero point ensuring proper requantization. Given a pixel array  $P$  (in this case, for either collateral or photometric pixel

<sup>1</sup>The requantization table is really two tables concatenated into one table with one section for the the SC data and one for the LC data, with buffers above, below, and between the LC and SC portions.

data), the first correction performed within CAL for the available rows ( $row$ ), columns ( $col$ ), and cadences ( $t$ ) is:<sup>2</sup>

$$P'_{all} (row, col, t) = P_{all} (row, col, t) - fixedOffset + meanBlack(t). \quad (5.5)$$

Note that MATLAB is used for all of the CAL science algorithms. To increase runtime performance, the operations are almost always performed on full or partial ( $nrows \times ncols \times ncadences$ ) pixel arrays as a whole rather than looping over any particular dimension. The ( $row, col, t$ ) notation here is to help the reader understand which pixels are processed in each step along with the dimensions of the pixel arrays and/or corrections. CAL only operates on the available (non-gapped) rows, columns, and cadences. The collateral pixels (black, masked smear, virtual smear, masked black, virtual black) need to be normalized by the number of spatial co-adds to convert to ADU/pixel/cadence. The original photometric pixels are already in these units.

$$\begin{aligned} P'_{black} (row, t) &= \frac{P_{black} (row, t)}{n_{black\ cols}} && (\text{LC + SC data}) \\ P'_{MS} (row, t) &= \frac{P_{MS} (row, t)}{n_{MS;rows}} && (\text{LC + SC data}) \\ P'_{VS} (row, t) &= \frac{P_{VS} (row, t)}{n_{VS;rows}} && (\text{LC + SC data}) \\ P'_{MB} (row, t) &= \frac{P_{MB} (row, t)}{n_{black\ cols} n_{MS;rows}} && (\text{SC data only}) \\ P'_{VB} (row, t) &= \frac{P_{VB} (row, t)}{n_{black\ cols} n_{VS;rows}} && (\text{SC data only}). \end{aligned} \quad (5.6)$$

#### 5.3.4 Black Correction

The black level (bias) in each CCD channel includes a voltage offset that has been added to the CCD output voltage to ensure that positive signals are input into the analog-to-digital converter (ADC). In practice, the bias voltage varies across the CCD readouts and has an intrinsic, low-level 2-D structure within each detector channel. The black correction must correct for the non-ideal behavior of the time- and spatially-varying black level.

The pipeline operator can select between one of three black correction modes:

- piece-wise static 2-D black followed by polynomial 1-D black model (*this method is available, but no longer used for processing*)
- piece-wise static 2-D black followed by a linear + exponential 1-D black model

<sup>2</sup>Note that throughout this chapter the use of the “prime” symbol on the variable on the left hand quantity of an equation indicates the new, corrected value of the quantity.

- dynamic time-dependent 2-D black model.

The selection is made based on the duration quarter being processed and the availability of the supporting data needed for the dynamic 2-D black model. Starting with SOC 9.2 the dynamic 2-D black model was the default selection except for the short quarters Q0, Q1, and Q17.

**5.3.4.1 Static 2-D black, Polynomial 1-D black** The channel-dependent 2-D black structures were measured for each channel during ground testing of the CCDs and are characterized in 2-D black models developed from unilluminated ground test images and reverse-clocked images (which ignore light hitting the CCDs) taken on the ground and during Commissioning. Some causes of the observed 2-D black structure include heating of the readout electronics, start of line (SOL) transients, and FGS frame and parallel transfer clocking crosstalk signals which are injected into the photometric signal as the image is read out. Figure 5.7 shows an example of a 2-D black model (top panel) which displays SOL features near the leading black region, and a close-up view (bottom panel) shows frame transfer (horizontal bands) and parallel transfer crosstalk (diagonal bands) signals. A 2-D black model (in ADU/exposure) is retrieved within CAL for each channel, scaled by the number of exposures, and is simply subtracted off all collateral and photometric pixels. Although time dependence of the 2-D black models is supported, in practice there is only one static model per channel available.

$$P'_{all}(row, col, t) = P_{all}(row, col, t) - 2Dblack(row, col, t). \quad (5.7)$$

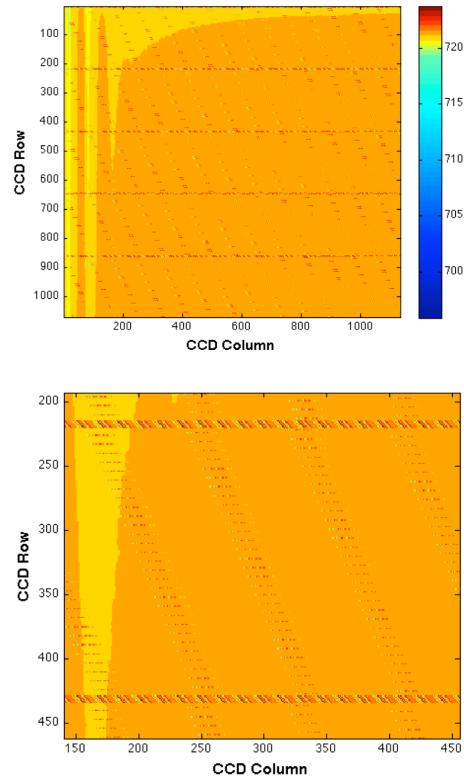
Once the 2-D black level is removed, a row dependent fit to the residual bias in the trailing black region is used to estimate a 1-D black correction. For software releases prior to SOC Release 7.0, a polynomial model (*polynomialOneDBlack*) was used with best fit model order determined in an iterative fashion using the Modified Akaike Information Criterion (Akaike, 1974). A robust fit was first performed to protect from outliers (neglecting charge injection rows in the virtual smear region), and a least squares with known covariance method was then used with the computed best polynomial order to produce the fit per cadence. This fitted polynomial is the black correction.

**5.3.4.2 Static 2-D black, Exponential 1-D black** In this mode, the same 2-D black described above is removed, followed by a row dependent fit to the residual bias in the trailing black region to estimate a 1-D black correction. The 1-D black model was developed to address issues in the previously used polynomial model caused by a flip-flopping polynomial order, which introduces chatter into the calibrated pixels. This empirical model (*exponentialOneDBlack*) is defined by six coefficients and includes a linear term in the masked smear region combined with two exponentials plus a linear term in the remainder of the rows. It has provided a much better fit to the 1-D black residuals than the previous polynomial model, resulting in a much improved black correction.

For each cadence, the black correction in a given row is subtracted from all available pixels in that row:

$$P'_{all}(row, t) = P_{all}(row, col, t) - p_{1-Dblack}(row, t). \quad (5.8)$$

For SC, the only collateral pixels downlinked are those that share a row or column with one of the target and background pixels. That is, only the SC target projections onto the trailing black rows and virtual/masked smear columns are available (see Figure 5.3). Since there are considerably fewer raw data to fit (1-5% compared to LC processing), the SC 1-D black correction is intrinsically noisier. Beginning with SOC 9.2, the LC 1-D black fit results which use the *exponentialOneDBlack* model were made available to SC CAL processing in order to improve the quality of the SC 1-D black fits. A bias term plus the LC 1-D black fit, linearly interpolated onto SC timestamps and scaled appropriately for the difference in cadence length, are fit to the available SC black pixels providing a more representative SC 1-D black correction. In order for SC CAL processing to take advantage of the LC results, LC 1-D black correction results using the *exponentialOneDBlack* model must be available and the black correction algorithm (module parameter `blackAlgorithm`) in SC processing must be set to *exponentialOneDBlack*.



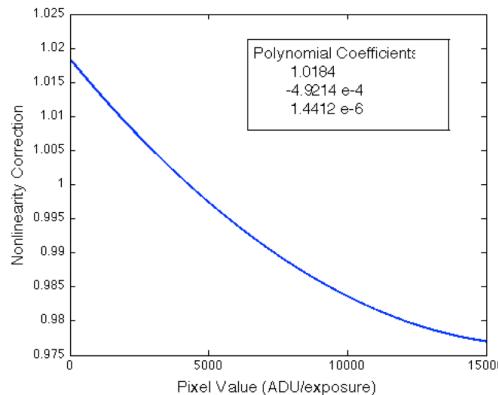
**Figure 5.7** An example of a 2-D black model (top, in units of ADU/exposure) and a close-up (bottom) that show the 2-D bias structure that is subtracted from all pixels. From Figure 5 of Quintana et al. (2010).

**5.3.4.3 Dynamic 2-D black** Time-dependent 2-D black models are available beginning with SOC 9.2. These models are produced per channel per cadence by DYNABLACK, a stand-alone pipeline module (see Chapter 4). They are available for both LC and SC processing (provided DYNABLACK has been run for that quarter) and are selected in CAL by setting the module parameter `blackAlgorithm = dynablack`. The `dynablack` correction is a direct substitution for the static 2-D black + 1-D black correction. The dynamic 2-D black goodness of fit is subjected to a number of tests within DYNABLACK and the fit residuals are too high, or too variable, on a given channel, CAL selects the static 2D-black + 1-D black models for use on that channel. The al-

gorithm that was actually applied on a given channel is recorded in the archive target pixel and light curve FITS files under the keyword BLKALGO (either *exponentialOneDBlack*, or *dynablack*).

Following the black correction (2-D black + 1-D black or dynablack), the black pixels are corrected for cosmic rays and saved for output to the Multi-mission Archive at Space Telescope (MAST) at STScI. Once the fits are performed, the pixels themselves are not used for any further calibrations. Note that CAL corrects only the collateral pixels for cosmic rays but uses the same methodology as the photometric pixel cosmic ray correction that is performed within PA (see Subsection 6.3.1).

### 5.3.5 Nonlinearity and Gain Correction



**Figure 5.8** The nonlinearity correction, shown for a sample CCD channel, is the fractional deviation from the linear electrons-to-ADU transfer function at each pixel value. From Figure 6 of Quintana et al. (2010).

The gain and nonlinearity describe the transfer function from photoelectrons ( $e^-$ ) in the CCD to ADU coming out of the ADC. Gain is the average slope of the transfer function, and ranges from 94 to 120  $e^-/\text{ADU}$  across the focal plane (Van Cleve & Caldwell, 2016; Caldwell et al., 2010). Nonlinearity is a measure of the relative deviation from a linear transfer function at each ADU signal level, expressed as a ratio. The nonlinearity is represented by a polynomial fit to the ratio as a function of input DN scaled to a single read. The nonlinearity model provides polynomial coefficients (Figure 5.8) that when evaluated, give the correction for the input black-corrected pixel values. The range of this correction across the focal plane is within  $\pm 3\%$  of linear. The nonlinearity model is valid up until the full-well level, which is the maximum number of electrons a pixel can hold before saturation occurs ( $\sim 10^6 e^-$ ). The gain model provides the gain value per channel and cadence in  $e^-/\text{ADU}$ , and all pixels are simply multiplied by the scalar gain following the nonlinearity correction. At this point,  $P_{all}$  now represents either smear arrays (in which the rows can be ignored since these are essentially just functions of columns and cadences) or photometric pixel arrays:

$$\begin{aligned} P'_{all}(row, col, t) &= P_{all}(row, col, t) \cdot p_{nonlin}(P_{all}(row, col, t)) \\ P'_{all}(row, col, t) &= P_{all}(row, col, t) \cdot G(t), \end{aligned} \quad (5.9)$$

where  $p_{nonlin}$  is the nonlinearity correction polynomial, and  $G(t)$  is the gain model over time.

### 5.3.6 LDE Overshoot/Uncorrected Correction

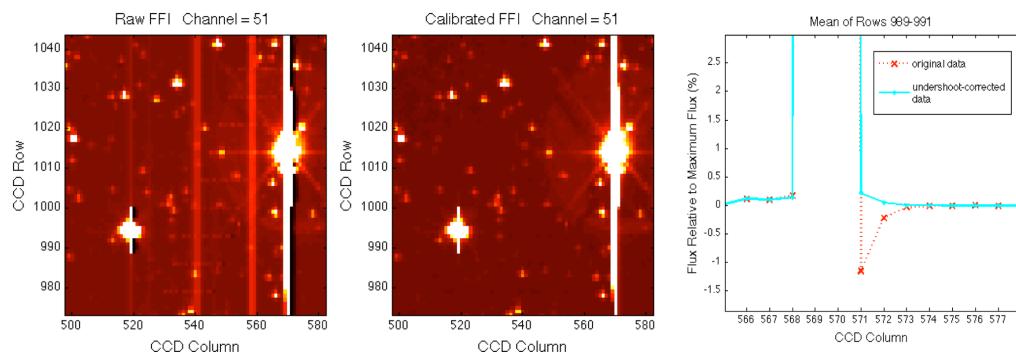
Overshoot and undershoot are signal distortions that were discovered during ground testing of the CCDs. They result from operating a clamp circuit in the local detector electronics (LDE) with insufficient bandwidth (Philbrick, 2009). The impulse response artifacts are most noticeable after light-to-dark (undershoot) and dark-to-light (overshoot) transitions, resulting in spikes in the pixel row time series of the affected targets (Figure 5.9). The undistorted image can be reconstructed by modeling these artifacts as a linear shift-invariant (LSI) system, which can be described by a set of difference equations that transforms an input signal  $x(n)$  into an output signal  $y(n)$ :

$$a(1)y(n) = b(1)x(n) + b(2)x(n-1) + \dots + b(n_b+1)x(n-n_b) - \quad (5.10)$$

$$a(2)y(n-1) - \dots - a(n_a+1)(n-n_a). \quad (5.11)$$

Here  $n-1$  is the filter order, and  $a$  and  $b$  are the feedback and feedforward filter coefficients, respectively, that determine the z-transform system response  $H(z)$ :

$$H(z) = \frac{b(1) + b(2)z^{-1} + \dots + b(n_b+1)z^{-n_b}}{a(1) + a(2)z^{-1} + \dots + a(n_a+1)z^{-n_a}}. \quad (5.12)$$



**Figure 5.9** A close-up stretched image of two saturated target stars that show pixel undershoot signatures resulting from bright-to-dark pixel transitions in the direction of the serial readout (left panel) and the calibrated image (middle). The mean of 3 pixel rows is shown for one target (right) with the undershoot response (the negative spike) along with the corrected pixel values. From Figure 7 of Quintana et al. (2010).

The undershoot model provides a set of 20 filter coefficients for  $a$ , and an inverse filter is applied (with  $b = 1$ ) to each row per cadence to correct for any

undershoot/overshoot:

$$P'_{all} (row, t) = \text{filter}(b, a, P_{all} (row, t)). \quad (5.13)$$

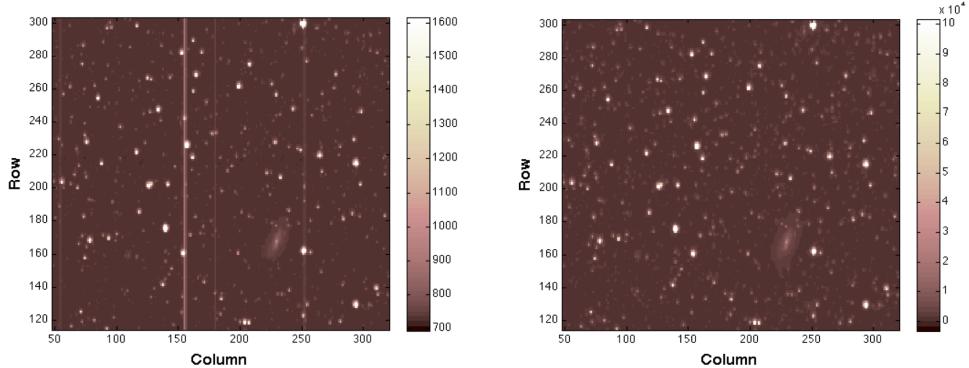
The function filter refers to the MATLAB built-in function based on the above difference equations. Note that an extra column of pixels on the side of the specified target pixels closer to the readout register is included in the target aperture to provide data for the undershoot/overshoot correction algorithm. Though it is designed to mitigate both undershoot and overshoot, we will refer to this filter below as simply the undershoot filter and to the effect as simply undershoot. Including this extra column, while certainly more helpful than not including it, is not sufficient to mitigate undershoot effects in cases where there is a bright star not downlinked for LC or SC processing that is within 20 columns upstream (to the readout register side) of a LC or SC target. The simple linear interpolation between target pixels along rows executed prior to applying the undershoot filter will mask undershoot due to these bright stars, resulting in up to a 3% error in the calibrated pixel flux for dim targets. Added functionality in SOC 9.2 allows estimation of these unseen bright stars in LC and SC processing using and average of the FFI data over the unit of work. This technique has been shown to reduce the error associated with applying the undershoot filter to below 0.1% for all targets. The use of FFIs to inform the undershoot filter applies to target and background pixels in both LC and SC processing. Also beginning with SOC 9.2, SC smear pixel processing is improved by providing the partially calibrated LC smear pixels per cadence to CAL SC processing to estimate and fill missing smear pixel values prior to applying the undershoot filter to the smear. Both the FFI inform and LC smear inform options have been enabled in the nominal pipeline configuration.

Both collateral and photometric data are corrected for undershoot. The median value of the correction across the focal plane array is  $\sim 0.34\%$  (Caldwell et al., 2010). In the collateral data invocation, the LC and SC masked and virtual smear pixels are next corrected for cosmic rays and saved for output. They are used to estimate the smear and dark current levels, which are then used to correct the photometric pixels (as described in the next section). The calibrated smear pixels exported to the MAST include CAL corrections up to this point (Thompson et al., 2016).

### 5.3.7 Smear and Dark Correction

The target and background pixels are corrected for both smear and dark current levels. Since the *Kepler* photometer is operated without a shutter, stars smear along columns as the CCD is read out, which is clearly visible in uncalibrated FFI data (Figure 5.10). Dark current is a thermally-induced signal in each physical pixel accumulated during an integration period, which includes the exposure time ( $t_{exp} \sim 6.02$  s) and readout time ( $t_{read} \sim 0.52$  s). Because the focal plane is maintained at such a cold temperature (-85 degrees C), the dark current is very

low with a median value of  $\sim 0.25\text{e}^- \text{ pixel}^{-1} \text{ sec}^{-1}$  across the focal plane. The smear and dark corrections are presented together because they are both estimated from linear combinations of the virtual and masked smear pixels.



**Figure 5.10** A portion of an uncalibrated FFI (ADU/cadence, left) and its calibrated image (photoelectrons per cadence, right) demonstrate the removal of smear from several columns. From Figure 8 of Quintana et al. (2010).

The masked smear pixels, which are shielded from star flux, detect dark current during integration ( $t_{exp} + t_{read}$ ) and collect the smear signal from the photometric and over-clocked virtual pixels during readout. The virtual pixels contain dark current that is accumulated during  $t_{read}$  only, but collect smear as they are clocked through the image. The dark level per cadence is computed by taking a robust mean of the masked and virtual smear differences –corrected for the different exposure times– from the common columns:

$$f_{dark} = \text{mean} \left\{ P_{MS}(col, t) - P_{VS}(col, t) \left( \frac{t_{exp} + t_{read}}{t_{exp}} \right) \right\}, \quad (5.14)$$

where  $f_{dark}$  is the dark current,  $P_{MS}$  is the estimated masked smear value in each column,  $P_{VS}$  is the estimated virtual smear value in each column,  $t_{exp}$  is the exposure time (6.54 sec) and  $t_{read}$  is the readout time (0.52 sec).

We interpolate dark level values over missing cadences to ensure that a dark level is available for all cadences. To compute the smear level, the dark level is first removed from the masked and virtual pixels:

$$\begin{aligned} P'_{MS}(col, t) &= [P_{MS}(col, t) - f_{dark}(t)] \\ P'_{VS}(col, t) &= \left[ P_{VS}(col, t) - f_{dark}(t)(col, t) \left( \frac{t_{read} + t_{read}}{t_{exp}} \right) \right]. \end{aligned} \quad (5.15)$$

Ideally, both masked and virtual pixels are available for each column and cadence, but either may be used if only one is available. If neither is available, however, the smear correction cannot be performed for that column. We use the ( $n_{cols} \times n_{cadences}$ ) logical gap indicator arrays  $G$  (where gaps = true) that are provided with the smear pixel arrays to track the missing data. The available smear pixels for each column are tracked using the logic shown in Table 5.3.

**Table 5.3** Smear estimate combination logic

Available Masked	Available Virtual	$C_{MS}$	$C_{VS}$
True	True	1/2	1/2
True	False	1	0
False	True	0	1
False	False	0	0

The available smear pixels for each column are tracked using the logic in the accompanying table, where  $C_{MS}$  and  $C_{MB}$  are coefficients in the linear combination of the dark-corrected masked and virtual smear pixels (where  $G'$  are logical arrays with gaps = false):

$$\begin{aligned} C_{MS}(col, t) &= \frac{1}{2}G_{MS}(col, t) \left[ 1 + G'_{VS}(col, t) \right] \\ C_{VS}(col, t) &= \frac{1}{2}G_{VS}(col, t) \left[ 1 + G'_{MS}(col, t) \right] \end{aligned} \quad (5.16)$$

and the smear estimate as a function of column and time is given by

$$f_{smear}(col, t) = P_{MS}(col, t) \cdot C_{MS}(col, t) + P_{VS}(col, t) \cdot C_{VS}(col, t). \quad (5.17)$$

The above smear and dark level estimates are computed during the collateral data calibration, resulting in a mean dark level per channel and an array of smear levels per column per channel. These are later subtracted from the photometric pixels in each column:

$$P'_{photo}(col, t) = P_{photo}(col, t) - f_{dark}(t) - f_{smear}(col, t). \quad (5.18)$$

An additional complication to the smear level estimate is bleeding charge from saturated targets into the masked or virtual smear regions, corrupting the smear estimate for that region. These effects are clearly visible in FFI data. Using a combination of FFI and raw LC collateral data, a map of which columns are corrupted by bleeding charge is developed for each quarterly observing season and made available to CAL pipeline processing. This map is channel- and season-dependent and is static throughout any one quarter. There are typically only one or two bleeding columns per channel. The map is used to gap either the virtual or masked smear estimate for the affected columns in the season of interest. See Thompson et al. (2016) §2.3.5.6 for a description of the format and use of the map.

### 5.3.8 Flat Field Correction

The flat field is the final major calibration step. It operates on photometric pixels to correct for spatial and temporal variations in pixel sensitivity to uniform illumination. Differences in pixel response can be due to variations in quantum

efficiency, throughput changes in the field flattener lenses, or anti-reflection coating of the CCD. The flat field model includes a geometric large-scale vignetting map combined with a small-scale (pixel-to-pixel) flat field map that is computed using a  $9 \times 9$  pixel high-pass filter (Van Cleve & Caldwell, 2016). The values represent the percent deviation from the local mean with a median value across the focal plane of  $\sim 0.96\%$ . The flat field model is applied to the photometric pixels for each cadence as shown below:

$$P'_{photo}(row, col, t) = \frac{P_{photo}(row, col, t)}{F_{flat}(row, col, t)}, \quad (5.19)$$

where  $F_{flat}$  is the flat field model. See Thompson et al. (2016) §2.3.5.11 and §2.3.5.12 for details of the models.

### 5.3.9 Additional Functionality in CAL

After the calibration is complete, the output fields are validated, fields are converted back to Java 0-based indices, and all information related to POU is saved. On the last CAL invocation, the achieved and theoretical compression efficiencies are computed. These metrics, along with time series of black, smear, and dark level metrics are computed within CAL and used by the photometer performance assessment (PPA) module to track and trend data.

The uncertainties are computed within CAL using the propagation of uncertainties (POU) algorithms (Clarke et al., 2010). The primary noise sources for *Kepler* include read noise (an additive noise source due to the readout process), quantization noise (stochastic, results from quantization in the ADC and pixel requantization) and Poisson-like shot noise. The uncertainties in the raw pixel data are computed at the start of CAL. If POU is disabled, the outputs from CAL are these raw pixel uncertainties corrected only for gain. This has been dubbed ‘minimal POU’. If POU is enabled, uncertainties are propagated at each transformation step by applying the transformation to the pixel covariance matrix. Since the full propagation of uncertainties is computationally intensive, the approach used in CAL is to compute the fully propagated uncertainties on a decimated set of cadences (typically 1 in 24 cadences is fully propagated) and then use the uncertainties obtained on this decimated set of cadences to compute the median difference from minimal POU. The uncertainties output from CAL with POU enabled are then minimal POU plus this median difference.

## 5.4 Summary

We have described the pixel-level corrections that are performed in the CAL pipeline for LC, SC, and FFI flight data. The data corrections include: 2-D and 1-D black, gain, nonlinearity, undershoot and overshoot distortions from the LDE electronics, cosmic rays, bleeding charge, dark current, smear, and flat field variations. The algorithms were validated using simulated flight data from

the End-To-End-Model (ETEM) (Bryson et al., 2010a) that was developed as a collaboration with Ball Aerospace Technologies Corporation. ETEM simulates every layer of the data – from CCD and instrument artifacts to transit light curves – and has proven to be a powerful tool in the development and testing of the pipeline modules. Output from CAL that is exported to the MAST includes raw and calibrated black, smear, and photometric pixels, along with the associated gap indicators and uncertainties. Additional metrics for cosmic ray detection, black level, smear level, and dark current level estimates are also provided.

## Bibliography

- Akaike, H., 1974. “A New Look at the Statistical Model Identification,” IEEE Transactions on Automatic Control, 19, 716
- Allen, C., Klaus, T., & Jenkins, J. 2010. “Kepler Mission’s Focal Plane Characterization Models Implementation,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401E–77401E–8
- Bryson, S. T., Jenkins, J. M., Peters, D. J., et al. 2010a. “The Kepler End-to-End Model: Creating High-Fidelity Simulations to Test Kepler Ground Processing,” in Proc. SPIE, Vol. 7738, Modeling, Systems Engineering, and Project Management for Astronomy IV, 773808
- Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010b. “Selecting Pixels for Kepler Downlink,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401D
- Caldwell, D. A., van Cleve, J. E., Jenkins, J. M., et al. 2010. “Kepler Instrument Performance: An In-Flight Update,” in Proc. SPIE, Vol. 7731, in Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave, 773117
- Clarke, B. D., Allen, C., Bryson, S. T., et al. 2010. “A Framework for Propagation of Uncertainties in the Kepler Data Analysis Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774020
- Haas, M. R., Batalha, N. M., Bryson, S. T., et al., 2010. “Kepler Science Operations,” ApJL, 713, L115
- Jenkins, J. M., & Dunnuck, J. 2011. “The Little Photometer that Could: Technical Challenges and Science Results from the Kepler Mission,” in Proc. SPIE, Vol. 8146, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 814602
- Klaus, T. C., McCauliff, S., Cote, M. T., et al. 2010. “Kepler Science Operations Center Pipeline Framework,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774017

- Philbrick, R. H. 2009. "Correction of Artifacts in Correlated Double-Sampled CCD Video Resulting from Insufficient Bandwidth," in Proc. SPIE, Vol. 7244, Real-Time Image and Video Processing 2009, 72440M
- Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010. "Pixel-Level Calibration in the Kepler Science Operations Center Pipeline," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401X
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)



## CHAPTER 6

---

# PHOTOMETRIC ANALYSIS: ALGORITHMS AND ARCHITECTURE

ROBERT L. MORRIS<sup>1</sup>, JOSEPH D. TWICKEN<sup>1</sup>, JEFFREY C. SMITH<sup>1</sup>, BRUCE D. CLARKE<sup>1</sup>,  
JON M. JENKINS<sup>2</sup>, STEPHEN T. BRYSON<sup>2</sup>, FORREST GIROUARD<sup>3</sup>, AND TODD C.  
KLAUS<sup>4</sup>

<sup>1</sup>The SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035, <sup>2</sup>NASA Ames Research Center, Moffett Field, CA 94035, <sup>3</sup>Logyx, LLC/NASA Ames Research Center, Moffett Field, CA 94035,  
<sup>4</sup>Stinger Ghaffarian Technologies, Inc./NASA Ames Research Center, Moffett Field, CA 94035

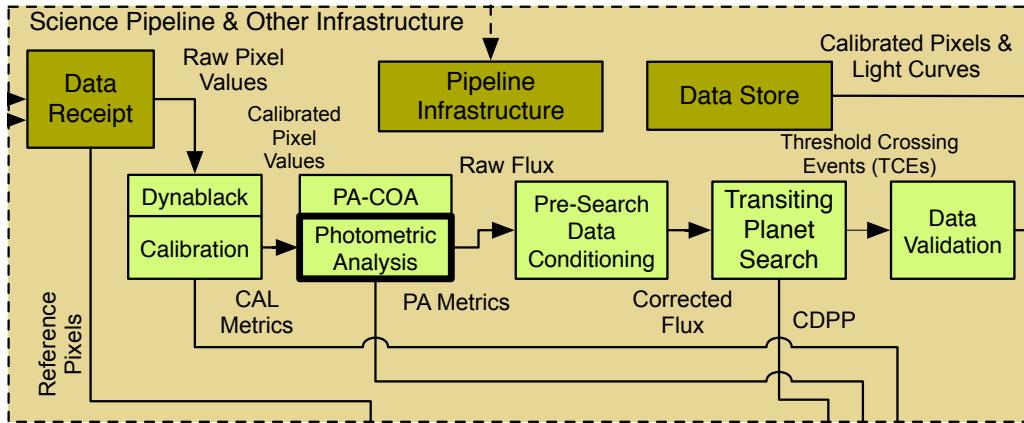
**Abstract.** With the *Kepler* mission in its closeout phase, the final results from the *Kepler* Science Operations Center (SOC) Science Processing Pipeline are now available for public use. We provide an overview of the architecture and algorithms of the Photometric Analysis (PA) pipeline component in its current and final state (SOC 9.3). It is our hope that this paper will assist interested parties in achieving a thorough understanding of the *Kepler* data products available from the Mikulski Archive for Space Telescopes (MAST). The primary functions of the PA module are to compute the photometric flux and photocenters (centroids) for over 165,000 long cadence ( $\sim$ thirty minute) and 512 short cadence ( $\sim$ one minute) stellar targets from the calibrated pixels in their respective apertures. We detail the main PA science algorithms for long and short cadences: cosmic ray cleaning, background estimation and removal, image motion estimation, optimal aperture selection, aperture photometry, and centroiding. Finally, we present examples of photometric apertures, raw flux light curves, and centroid time series from *Kepler* flight data.

**Keywords:** *Kepler*, transit photometry, light curve, raw flux, photocenter, centroid, spectroscopy.

### 6.1 Introduction

The Photometric Analysis (PA) module of the *Kepler* Science Data Processing Pipeline is responsible for generating the flux time series and centroid time series for each target star observed by *Kepler*. PA processes data from both long cadence (LC) target stars, which are sampled at 29.4 min, and short cadence (SC) target stars, which are sampled at 58.8 s (Jenkins et al., 2010b). PA operates on the calibrated pixels produced by the calibration (CAL) pipeline module prior to the identification and removal of instrumental signatures and systematic errors by the Presearch Data Conditioning (PDC) pipeline module in preparation for the transit search. Figure 6.1 shows the PA module in the context of the Science Operations Center (SOC). The primary function of the PA module is to compute the photometric flux and photocenters (centroids) for stellar targets from

the calibrated pixels in their respective apertures. Prior to computation of the photometric flux for each target, so-called Argabrightening events are mitigated (Witteborn et al., 2011; Christiansen et al., 2013, Section 5.8), cosmic rays are removed, and a background estimate is subtracted from the pixels in the target apertures. A subset of the pixels collected for each target is then selected to optimize a photometric figure of merit. The photometric flux at each cadence is calculated from these optimal apertures by summing the constituent pixels, a process known as simple aperture photometry (SAP). Secondary functions of PA are to compute metrics to monitor instrument performance and to support systematic error correction in the PDC module. Raw flux light curves and centroid time series are exported to the Mikulski Archive for Space Telescopes (MAST) (see Thompson et al., 2016, Section 2.3). Corrected flux light curves generated in PDC are also exported to the MAST. These account for systematic errors due to data and spacecraft anomalies, and also for excess flux in the target apertures due to background sources.



**Figure 6.1** Photometric Analysis (PA) in the context of the architecture of the SOC. PA identifies and removes cosmic rays from calibrated background and target star pixels produced by the calibration (CAL) component, fits and removes sky background flux from the target star pixels, and then estimates the brightness and position of each target star in each cadence. The flux time series produced by PA are then conditioned prior to the transit search in Presearch Data Conditioning (PDC) to remove instrumental signatures and residual outliers.

PA has undergone continuous development and improvement since its inception, with an emphasis on boosting pipeline sensitivity to transit events and increasing throughput. Major improvements to PA since its first application in the *Kepler* Science Data Processing Pipeline (Twicken et al., 2010b) include the addition of internal optimal aperture computations (previously done externally without the benefit of analysis of pixel data<sup>1</sup>), an improved cosmic ray cleaning algorithm, and parallelization of target processing. Minor improvements include

<sup>1</sup>The previous version of the optimal aperture selection algorithm made use of the measured image motion over the course of the observations within a quarter but relied on focal plane models and the stellar catalog to perform the signal-to-noise ratio (SNR) calculation, yielding the optimal aperture. See Chapter 7 for more details.

identification of cadences during which reaction wheel speeds dropped to or near zero and computation of contamination levels in optimal apertures by so-called *rolling band* image artifacts. An overview of the PA module's architecture and component algorithms is presented in Section 6.2. PA science algorithms are detailed in Section 6.3, including cosmic ray cleaning, background estimation and removal, image motion estimation, optimal aperture selection, aperture photometry, and target centroiding. Conclusions are presented in Section 6.4.

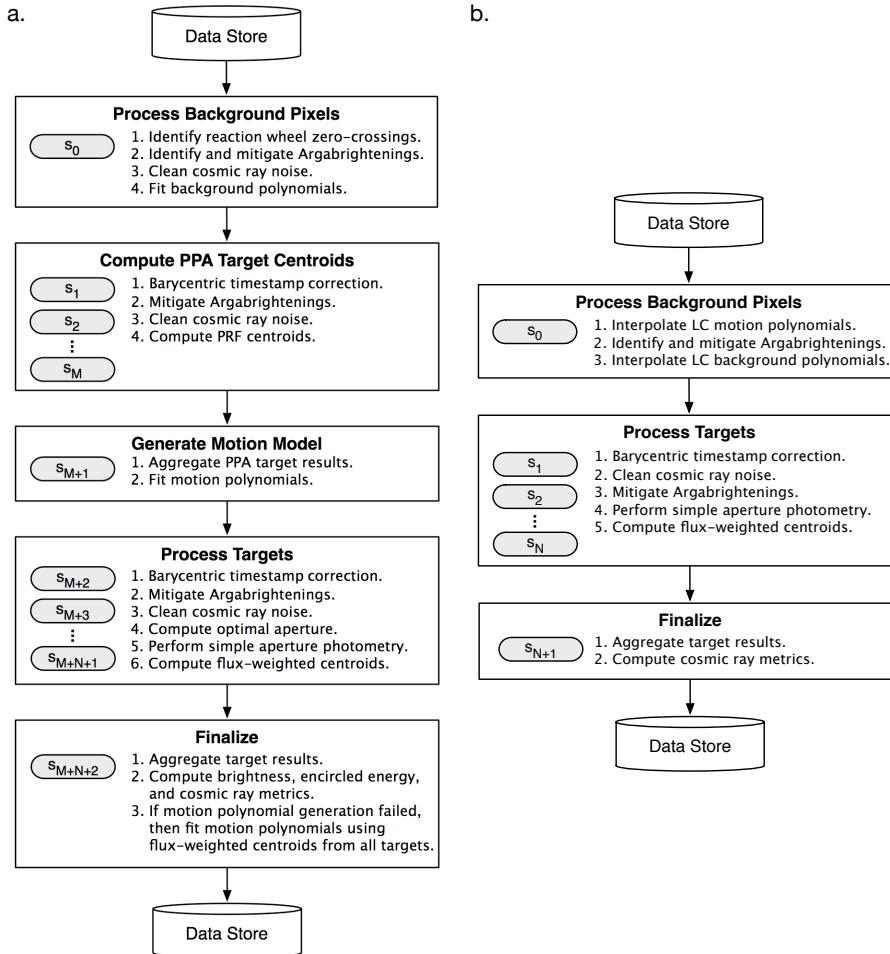
## 6.2 Architecture

The standard PA *unit of work* (Klaus et al., 2010a,b) for LC and SC science data processing is a single module output for a duration of one quarter (LC) or one month (SC – Twicken et al., 2010b). Science data processed by PA include calibrated background and target pixels. Targets may be stars, in which case PA selects a photometric aperture automatically, or they may be manually-specified custom apertures. A subset of  $\sim 200$  stellar targets on each channel is used by PA to derive a set of high-quality centroids to which a polynomial model of image motion is fitted. These Photometer Performance Assessment (PPA) target sets were initially preselected for use in attitude reconstruction by the PPA pipeline module (Jenkins et al., 2010b; Li et al., 2010) and possess the following characteristics that make them useful elsewhere in the pipeline: bright but not saturated, uncrowded, and well-distributed across each module output.

By definition a PA *task* processes a single unit of work. As illustrated in Figure 6.2, a LC task consists of the following five sequential steps: 1) background pixel processing, 2) PPA target centroiding, 3) motion model estimation, 4) target processing, and 5) results aggregation. Each step is executed as one or more *subtasks*. As shown in Figure 6.2, the target processing and PPA centroiding steps are divided into multiple subtasks and can be performed in parallel if the computational environment and resources allow. Execution time can be minimized by processing each target in a separate subtask, though in practice limited resources have necessitated multiple targets per subtask. Since there are neither background pixels nor full sets of PPA targets in SC units of work, processing relies on interpolated versions of the LC background and motion models.

The major algorithmic components comprising each subtask are introduced here and described in greater detail in Section 6.3:

**Identify Reaction Wheel Zero-Crossing Events:** Static friction causes the space-craft's reaction wheels to exhibit a low-amplitude rumble at wheel speeds near zero, which can have a small but measurable effect on pointing and photometry. Ancillary engineering data, including wheel speed measurements, are delivered along with data from the photometer and are analyzed to determine and flag cadences on which any wheels speeds approached zero. First, a median filter is applied to the wheel speed time series for all four wheels and strings of consec-



**Figure 6.2** Flow diagrams for LC and SC PA tasks, which each process a single unit of work. Panel a) Each long-cadence PA task is divided into  $M + N + 3$  subtasks, represented by shaded ovals. Subtasks within the PPA centroiding and target processing steps are executed in parallel if the computational environment allows; otherwise, subtasks  $s_0$  through  $s_{M+N+2}$  are executed sequentially. The algorithmic steps performed by each subtask are listed on the right-hand side of the box containing it. Panel b) There are no background pixels in SC data, and SC processing relies on interpolated versions of the LC background and motion models.

utive zeros in the filtered data are identified. The median filter length is configurable and is set to 5 for the SOC 9.3 run. Then strings longer than the median filter length are flagged as zero-crossing events. Since the engineering data are not synchronized with the photometer data (they are sampled at a frequency of approximately 0.0083 Hz), the final step is to determine and flag cadences that overlap with the flagged zero-crossing events in the engineering data.

**Identify and Mitigate Argabrightening Events:** It has been observed in flight data that there are cadences for which the flux values of all background and

target pixels are elevated (Jenkins et al., 2010b). These have been dubbed *Argabrightening* events after Vic Argabright, the Ball Aerospace engineer who first observed and reported them. There has been speculation about the source of the Argabrightening events, but the true source remains unknown. If Argabrightening mitigation is enabled (the default), then detection statistics are formulated for each cadence and a threshold is applied to identify the Argabrightening cadences. Identification of Argabrightening cadences is done independently for each unit of work and occurs in the first subtask of both LC and SC tasks. Argabrightening cadences are subsequently gapped in all background and target pixel time series in the unit of work. Argabrightening statistics are formulated from the background pixels for LC units of work and from target pixels outside of the optimal apertures of the respective targets in the first subtask for SC units of work (Twicken et al., 2010b).

**Clean Cosmic Ray Noise:** If cosmic ray cleaning is enabled (the default), the effects of cosmic ray strikes are identified and removed from the calibrated background (LC only) and from target pixels (both LC and SC). Lists of detected cosmic ray events and the corrections applied are maintained for both background and target pixels and are written into the pipeline’s Data Store in the final subtask.

**Fit Background Polynomials:** For LC units of work, a 2-D (background) polynomial is fitted as a function of CCD row and column to the cosmic ray-cleaned background pixels on a cadence-by-cadence basis.

**Fit Motion Polynomials:** In LC units of work, a polynomial model of image motion is constructed by fitting a pair of 2-D polynomials to the collection of PPA target centroids at each cadence: one mapping celestial coordinates (Right Ascension (RA) and Declination (Dec)) to CCD row coordinates and another mapping them to column coordinates. In SC units of work, motion polynomials are obtained by interpolating LC motion polynomials retrieved from the Data Store. It follows that LC processing precedes that of SC. Within PA, motion polynomials are used to account for image motion in cosmic ray cleaning and in image modeling for optimal aperture selection. They are also utilized elsewhere in the pipeline for focal plane geometry fitting (Tenenbaum & Jenkins, 2010), systematic error correction (Twicken et al., 2010a), attitude determination (Li et al., 2010), and computation of instrument performance metrics (Li et al., 2010).

**Optimal Aperture Selection:** PA reexamines the optimal apertures provided by the Target and Aperture Definitions (TAD) component based on the observed image motion (see Chapter 3 and Chapter 7 for more details). Since TAD computes optimal apertures without access to the observed pixels, there is an opportunity for PA to use pixel data to improve the apertures. For each target mask, PA selects a photometric aperture based on a number of factors, including the

estimated SNR at each pixel, the estimated Combined Differential Photometric Precision (CDPP) of light curves produced by candidate apertures, and several tuned heuristics.<sup>2</sup>

**Simple Aperture Photometry:** PA light curves are computed by summing the cosmic ray-cleaned and background-subtracted pixels within the optimal photometric aperture. For SC units of work, background polynomials must be normalized and interpolated at the SC timestamps before they can be used to obtain background estimates. Simple Aperture Photometry (SAP) is the only photometry method supported by PA.

**Compute Target Centroids:** Row and column centroid time series are computed from the background-removed target pixels on a cadence by cadence basis. Flux-weighted centroids are computed for all PA targets. Centroids may also be computed by fitting a predetermined Pixel Response Function (PRF – Bryson et al., 2010a) to calibrated target pixel values. In practice, PRF centroiding was only done for PPA targets. It was disabled for non-PPA targets both because of the computational expense and because the quality of results diminishes for dimmer and more crowded stars.

**Compute Encircled Energy Metric, Brightness Metric, and Cosmic Ray Metrics:** The encircled energy metric is computed for each cadence from the background-removed target pixels and centroids of PPA targets. This metric is a robust measure of the average radius required to capture a specified fraction of flux from the respective targets. The brightness metric is computed for each cadence from the raw flux light curves for the set of PPA targets and is a robust measure of the average ratio of observed-to-estimated flux for the given targets.

Cosmic ray metrics are computed separately from the background and target cosmic ray event lists. The metrics are computed for each cadence in a unit of work and include the number of events per square centimeter per second detected during the cadence, along with the first four central moments (mean, variance, skewness, and kurtosis) of the set of event magnitudes.

### 6.3 Photometric Analysis Science Algorithms

The PA module’s constituent algorithms were introduced in the previous section. In this section we provide detailed descriptions of the algorithms for cosmic ray cleaning, background and motion polynomial model fitting, centroiding, optimal aperture selection, and simple aperture photometry. The uncertainties associated with calibrated pixels presented to PA are propagated by these algorithms to produce uncertainties for each output flux and centroid value. Due to the large

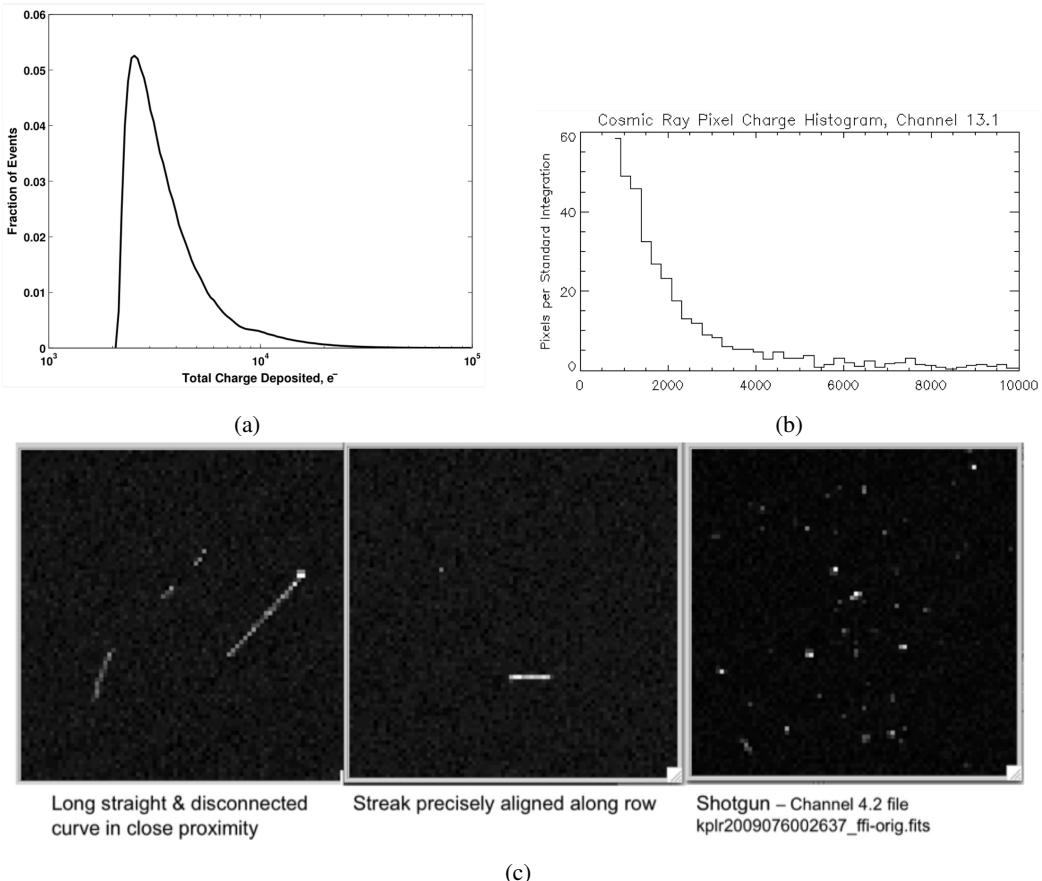
<sup>2</sup>CDPP is a measure of the effective noise seen by a transit of a given duration. A 10 ppm CDPP at 6 hours indicates that a 10 ppm deep 6-hour transit is a 1- $\sigma$  event (Jenkins et al., 2010c; Christiansen et al., 2012).

computational expense, rigorous propagation of uncertainties on every cadence was not possible. In SOC 9.3, uncertainties for every cadence were propagated under an assumption of statistical independence between variables (i.e., using diagonal covariance matrices). A more rigorous treatment, utilizing full covariance matrices for a subset of cadences to estimate and correct the bias of the approximations (Clarke et al., 2010; Twicken et al., 2010b), was included as an option but was not used in the final processing. The resulting uncertainties are likely understated by a few tenths of a percent to several percent over the range of target star magnitudes  $Kp=9$  to  $Kp=16$ .

### 6.3.1 Cosmic Ray Cleaning

Galactic cosmic rays and solar energetic particles (we loosely refer to all such particles as “cosmic rays” regardless of their origin) continuously bombard the instrument with an expected mean flux of  $\sim 5 \text{ cm}^{-2} \cdot \text{s}^{-1}$  (Van Cleve & Caldwell, 2016). When a cosmic ray strikes the detector, it typically produces free electrons in more than one pixel, depending on its energy and angle of incidence. We refer to the charge deposited in an individual pixel as a *pixel event* and the set of all pixel events caused by a cosmic ray as a *particle event*. Based on simulations, approximately 18% of pixels are expected to suffer pixel events on a given LC (0.6% for SC). The total charge released in a particle event is expected to follow the distribution shown in Figure 6.3a, which was obtained by simulation (Jenkins et al., 2004). Since the photometer has no shutter, cosmic ray noise characteristics cannot be estimated from dark images during mission operations. However, during *Kepler*’s commissioning there was a unique opportunity to measure the cosmic ray signal in the absence of stellar illumination before the instrument’s dust cover was jettisoned. Figure 6.3b illustrates the observations from this period (Van Cleve & Caldwell, 2016), which generally agree with our expectations.

The effects of cosmic ray strikes appear as positive single-cadence impulses in a pixel time series. PA cleans (i.e. identifies and removes) cosmic ray effects from both background and target pixels immediately after Argabrightening events have been mitigated. Cosmic ray identification in the relatively low flux background pixels is more effective than in target pixels where deposited energy is more often a small fraction of the total flux. Cosmic ray strikes are expected to deposit less than  $420 e^-$  in roughly 90% of all affected pixels during a LC measurement, while the shot noise-dominated background levels are on the order of  $10^5 e^-$ . The vast majority of cosmic ray effects are therefore below the background noise level and cannot be reliably separated from the other signal components. The problem of reliably detecting even large effects is nontrivial because high-frequency stellar variations and image motion can also produce significant impulse-like features in pixel flux time series. There is also the practical challenge of cleaning approximately 5,000,000 pixel time series with limited

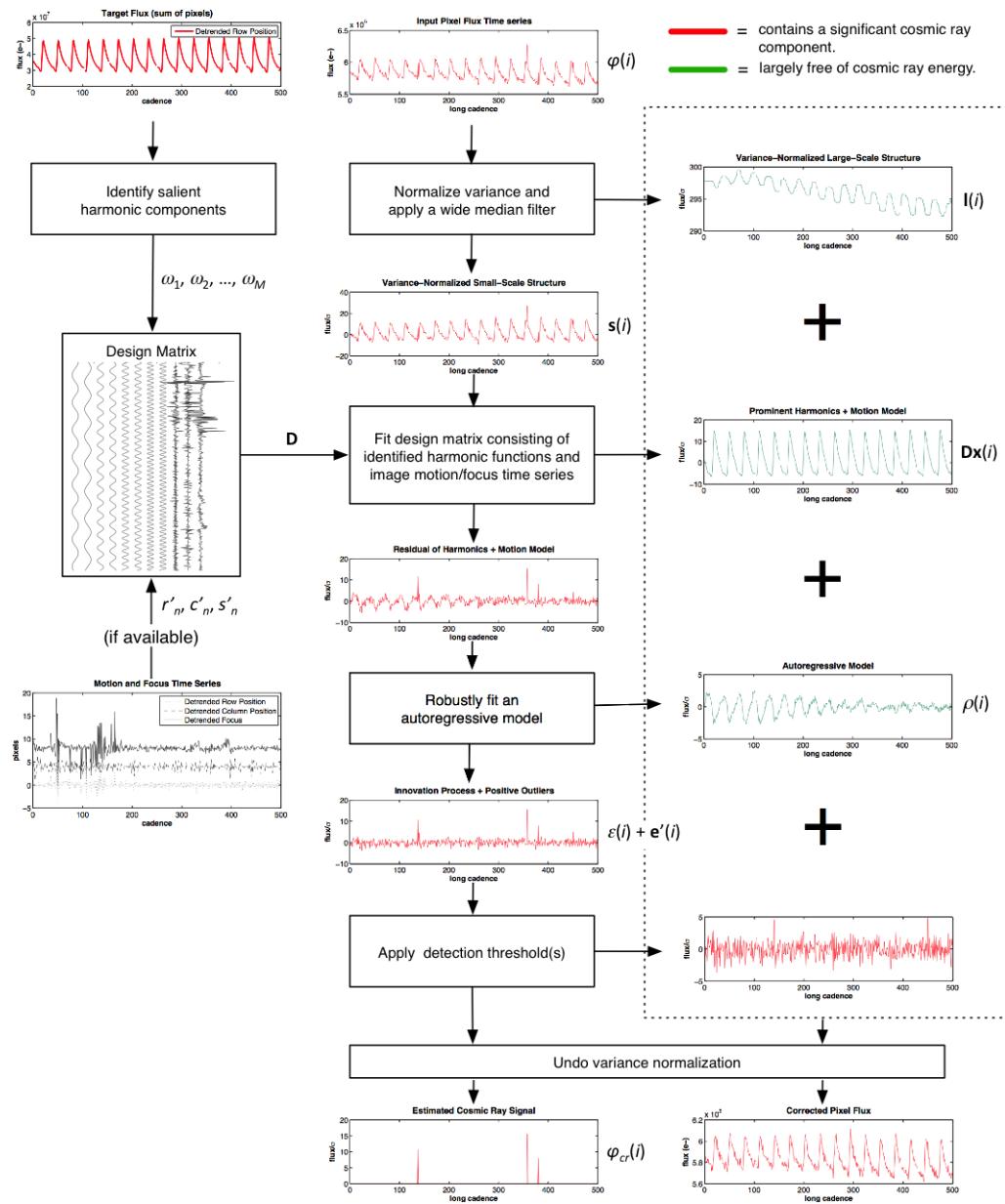


**Figure 6.3** Characteristics of cosmic ray noise. a) Estimated distribution of photoelectrons per particle event (Jenkins et al., 2004). b) Histogram of observed charge deposited per pixel event, derived from channel 13.1 full-frame images. Only pixel events exceeding a  $6\sigma$  read noise threshold are shown. c) Examples of cosmic ray strikes before ejection of the instrument’s dust cover (Van Cleve & Caldwell, 2016).

time and computational resources, which limits the allowable complexity of a solution.

Our algorithm attempts to quickly isolate and remove large-amplitude cosmic ray effects from the pixels in each target mask (this applies both to 4-pixel background targets and stellar targets). As illustrated in Figure 6.4, our strategy is to successively model and remove features of decreasing scale from the time series until only white noise with single cadence outliers remains. At that point we apply a threshold to clean likely cosmic ray effects. In the interest of speed we fill short gaps by linear interpolation. Short gaps are defined by a configurable threshold (typically 10 cadences). Rather than attempting to fill longer gaps, we process pixel time series in segments, the boundaries of which are determined by the long gaps.

The first step is to normalize the variance of the observed pixel time series by dividing each calibrated flux measurement  $\phi(i, n)$  by its corresponding uncertainty value  $\sigma(i, n)$ , where  $i$  denotes a particular pixel and  $n$  a cadence, to pro-



**Figure 6.4** Flow diagram of the cosmic ray cleaning algorithm showing the decomposition of a pixel time series in order to isolate large-amplitude cosmic ray effects. Flux components plotted in red contain significant cosmic ray-induced noise, while green components should be largely free of cosmic ray energy. Note that in the detection step it is impossible to distinguish low-level cosmic ray noise from fluctuations in the innovation process of the AR model. Our goal is to remove the large outliers.

duce a variance-normalized time series. We then model the variance-normalized, calibrated flux time series  $\mathbf{f}_i$  at pixel  $i$  as

$$\hat{\mathbf{f}}(i) = \mathbf{l}(i) + \mathbf{D}\mathbf{x}(i) + \mathbf{a}(i), \quad (6.1)$$

where  $\mathbf{l}(i)$  denotes the large-scale trend in the time series,  $\mathbf{D}\mathbf{x}(i)$  denotes a model that accounts for both salient harmonic components and the effects of image motion, and  $\mathbf{a}(i)$  is an autoregressive (AR) process. The AR model can be further decomposed into the sum of a white *innovation process* (Papoulis, 1986)  $\epsilon(i)$  and a prediction  $\rho(i)$  derived from it (Equation 6.3). Provided the model adequately captures the effects of image motion and stellar variability, cosmic ray noise will be confined to  $\epsilon(i) + \mathbf{e}(i)$ , where  $\mathbf{e}(i) = \mathbf{f}(i) - \hat{\mathbf{f}}(i)$  denotes the model residual. Once the model has been fit, as described below, we identify and correct positive impulses in  $\epsilon(i) + \mathbf{e}(i)$  that are above a configurable detection threshold (typically  $4\sigma$ ).

In constructing the model of Equation 6.1, we separate the structure of the variance-normalized pixel time series into large- and small-scale components  $\mathbf{l}(i)$  and  $\mathbf{s}(i)$ , respectively, by applying a 49-cadence median filter. At this point, cosmic ray-induced impulses should be isolated in the small-scale component. It is important to be aware that cosmic rays occasionally produce significant effects on multiple cadences by causing changes in pixel sensitivity, termed Sudden Pixel Sensitivity Dropouts (SPSD – Christiansen et al., 2013, Section 5.9), and that such events are corrected separately in the PDC pipeline module (Stumpe et al., 2012).

Next we model and remove the effects of periodic stellar variations and abrupt changes in spacecraft pointing and/or instrument focus, both of which can produce impulse-like features that resemble cosmic ray effects. For a given target mask consisting of pixels with indices in the set  $I$ , the summed small-scale component  $\sum_{i \in I} \mathbf{s}(i)$  is analyzed to identify up to  $M_{max}$  (a configurable parameter typically set to 25) salient harmonic components with frequencies  $w_1, w_2, \dots, w_M$ . If motion polynomials are available, they are used to produce estimates of centroid position and plate scale (a convenient proxy for the system’s state of focus) at each cadence (see Appendix 6-A for details). The position and focus time series are then detrended with a 49-cadence median filter to isolate short-timescale changes in pointing and focus. Note that motion polynomials are typically available for stellar target processing, where they are most useful, but not for background target processing, as background targets don’t have celestial coordinates associated with them.

The design matrix,

$$\mathbf{D} = \left[ \begin{array}{ccc|ccc} h_{1,1} & \dots & h_{1,M} & r'_1 & c'_1 & s'_1 \\ h_{2,1} & \dots & h_{2,M} & r'_2 & c'_2 & s'_2 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ h_{N,1} & \dots & h_{N,M} & r'_N & c'_N & s'_N \end{array} \right], \quad (6.2)$$

is constructed from the identified harmonics and the detrended motion time series, where  $M$  is the number of salient harmonics identified and  $N$  is the number of cadences in the unit of work being processed. The coefficients  $h_{n,m} = \cos(w_m t_n) \sin(w_m t_n)$  in the  $M$  columns to the left of the dashed line model the harmonic content, where  $t_n$  is the mid-cadence timestamp of the  $n^{th}$  cadence, while columns to the right model effects of image motion. Terms  $r'_n$ ,  $c'_n$ , and  $s'_n$  denote median-filtered estimates at cadence  $n$  of the row and column centroid positions,  $r$  and  $c$ , and plate scale  $s$ , respectively. We then solve  $\mathbf{s}(i) = \mathbf{D}\mathbf{x}(i)$  for coefficients  $\mathbf{x}(i)$  in the least squares sense by computing the pseudoinverse of  $\mathbf{D}$ . Singular value decomposition is used to obtain a solution in cases where  $\mathbf{D}$  is poorly conditioned or has low rank; otherwise a QR decomposition is used.

After subtracting the large scale trend and the harmonic/motion model, the remaining flux  $\mathbf{s}(i) - \mathbf{D}\mathbf{x}(i)$  at pixel  $i$  is modeled as a zero-mean AR process  $\mathbf{a}(i)$  with infrequent outliers. Given parameters  $b_1(i), \dots, b_{K_{AR}}(i)$ , the  $n^{th}$  element of  $\mathbf{a}(i)$  is given by

$$a(i, n) = \rho(i, n) + \epsilon(i, n) = \sum_{k=1}^{K_{AR}} b_k(i) a(i, n-k) + \epsilon(i, n), \quad (6.3)$$

where  $\rho(i, n)$  can be thought of as a prediction of  $a(i, n)$  based on the preceding  $K_{AR}$  values in the time series and  $\epsilon(i, n)$  as the prediction error. The model order  $K_{AR}$  is a configurable parameter. If the time series being modeled is wide-sense stationary (having shift-invariant mean and autocovariance), then the innovation process  $\epsilon(i)$  should be white noise. We use the Burg method (Burg, 1967) to robustly fit the AR parameters to the remaining flux in two iterations: a forward pass and a backward pass. On the forward pass we identify and flag outliers exceeding a threshold (typically  $10\sigma$ ), which are replaced by the innovation's mean value (zero). Upon completion of the backward pass, we once again identify and zero any outliers and then compute the residual  $\mathbf{e}(i)$  of Equation 6.1. At this point the bulk of cosmic ray effects should be confined to  $\epsilon(i) + \mathbf{e}(i)$ .

The final step in our algorithm consists of analyzing  $\epsilon(i) + \mathbf{e}(i)$  to identify positive-going impulses above the detection threshold  $\tau$  (typically  $4\sigma$ ). We start by applying a 3-cadence median filter to remove any non-impulsive features from the residual. The result,  $\mathbf{e}'(i)$ , is then added to the estimated innovation from Equation 6.3, which also contains significant cosmic ray noise, and the detection threshold is applied to obtain an estimate of  $\phi_{cr}(i)$ , the large-amplitude cosmic ray signal:

$$\hat{\phi}_{cr}(i, n) = \begin{cases} (\epsilon_{i,n} + e'(i, n))\sigma(i, n), & \text{if } \epsilon(i, n) + e'(i, n) > \tau \\ 0, & \text{otherwise} \end{cases}. \quad (6.4)$$

Note the multiplication by  $\sigma(i, n)$  in Equation 6.4 that undoes the initial variance normalization. To apply the correction for pixel  $i$  we simply subtract the estimated cosmic ray signal  $\hat{\phi}_{cr}(i)$  from the calibrated pixel flux  $\phi(i)$ .

### 6.3.2 Background Estimation and Removal

Background targets are acquired at the LC rate in a quasi-grid pattern on each of the focal plane module outputs (see Subsection 3.2.3). 1,116  $2 \times 2$  pixel apertures are defined on each channel. The background pixel time series represent spatial and temporal samples of the global background level. In selecting background targets an effort is made to prevent them from being corrupted by flux from neighboring stars or from saturated targets on the same columns. Nevertheless, there are background pixels in the flight data that exhibit high flux levels for reasons including nearby or saturated targets and imperfectly corrected smear. These pixels necessitate the development of a method for background estimation and subtraction that is robust against outliers.

Background estimation is performed in two steps. The process begins by fitting a 2-D (background) polynomial to the calibrated, cosmic ray-corrected background pixel values as a function of the CCD row and column coordinates for each cadence. This occurs in the first PA subtask in each LC unit of work. In all subsequent target subtasks, the background is estimated by evaluating the background polynomial for each cadence at the CCD coordinates of the respective target pixels. This produces a spatially smooth estimate of the local background for each target without dedicating any background pixels to specific targets. The ratio of background pixels to pixels from stellar or custom apertures is on the order of 1:14 across the entire focal plane.

Once the background is estimated for each target pixel and cadence, it is removed by subtracting the estimated value from the calibrated, cosmic ray-corrected target pixel value. Uncertainties in the background-removed target pixels are propagated as described at the beginning of this section.

Background removal for SC targets is complicated by the fact that there are no SC background pixels. In SC units of work, the LC background polynomials are provided as input to PA in the first subtask. The background polynomials are then scaled to account for the shorter integration times and interpolated in time at the midpoints of the SC intervals. The background estimation and removal process proceeds with the interpolated background polynomials as before. It should be noted that changes in the background that occur on time scales shorter than the LC rate cannot be captured in the SC background estimation process.

The order for the 2-D polynomial that is fit to the background pixels for each LC is determined in the pipeline with the Akaike Information Criterion (AIC – Akaike, 1974). This criterion includes a penalty that increases with fit order and seeks to optimize the trade-off between fit order and goodness of fit. In PA, the order used in the background fit for all cadences is determined by the order (up to a specified maximum) that minimizes average AIC over all cadences in the unit of work. For a background fit of order  $K$ , the number of background polynomial coefficients for each cadence is  $(K + 1)(K + 2)/2$ .

For the set of calibrated, cosmic ray-subtracted background pixels, let  $\phi_b(i)$  and  $\sigma_b(i)$  designate, respectively, the pixel values and uncertainties of pixel  $i$

for a given LC. Furthermore, let the (one-based) CCD row and column coordinates at the center of pixel  $i$  be designated by  $r_0(i)$  and  $c_0(i)$ , respectively. The background estimate for pixel  $i$  is given by

$$\hat{\phi}_b(i) = \sum_{k=0}^{K_B} \sum_{l=0}^k B_q r_0^{k-l}(i) c_0^l(i), \quad (6.5)$$

where the background polynomial coefficient index is given by

$$q = k(k + 1)/2 + l. \quad (6.6)$$

Letting  $I$  denote the set of background pixel indices, the background polynomial coefficients

$B_0, B_1, \dots, B_{K_B}$  for the cadence in question are determined by minimizing the weighted chi-square,

$$\chi_b^2 = \sum_{i \in I} \left( \frac{\phi_b(i) - \hat{\phi}_b(i)}{\sigma_b(i)} \right)^2. \quad (6.7)$$

For the set of pixels  $A$  in a specified target aperture, the background level may then be estimated for the given cadence from the background polynomial  $B$  and subtracted from the calibrated, cosmic ray-corrected target pixels to obtain the background-removed target pixel values  $\phi_s$  as follows:

$$\phi_s(i) = \phi(i) - \hat{\phi}_{cr}(i) - \hat{\phi}_b(i). \quad (6.8)$$

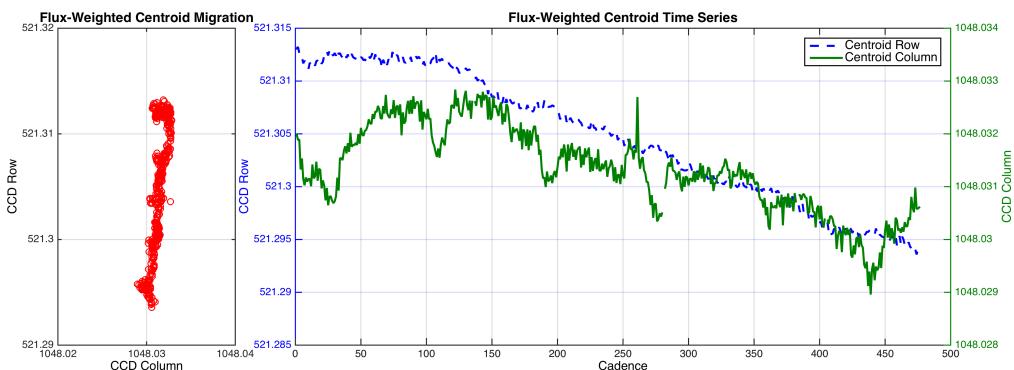
### 6.3.3 Centroiding

There is substantial motion of the target positions on the focal plane in the flight science data at the sub-pixel level. The dominant source of long-term target motion is differential velocity aberration (DVA), which causes the targets to trace small but significant elliptical paths across the respective CCD detectors over the period of the heliocentric orbit of the photometer. The maximum motion due to DVA is 0.6 pixels per observing quarter (Jenkins et al., 2010a). There is also target movement due to pointing jitter, pointing drift, focus changes (due mainly to temperature changes in the photometer), and commanded attitude adjustments to compensate for pointing drift. It should be noted that the photocenters of variable targets also move in crowded apertures.

Flux levels vary with target motion at even the sub-pixel level. Systematic effects that result in target motion therefore produce correlated signatures in the associated light curves. Precise computation of the target locations on each cadence is critical for correcting systematic errors later in the pipeline. It is also required for precise *a posteriori* definition of optimal target apertures (see Sub-section 6.3.5), for precision reconstruction of spacecraft attitude, and for monitoring instrument performance. The photocenter of each target in its aperture is

referred to as the target *centroid*. It should also be noted that analysis of centroid motion (Wu et al., 2010) is a critical tool for distinguishing between legitimate transits due to orbiting planets and apparent transits due to background eclipsing binaries or transits on background stars.

Two centroiding methods are employed in PA. Flux-weighted centroids are computed for every target and cadence and are currently exported to the MAST. These centroids are computationally inexpensive and essentially determine the photometric center of mass in the target aperture. Subsection 6.3.3 illustrates the sub-pixel migration of an 11<sup>th</sup> magnitude star's flux-weighted centroid during Q0. PRF-based centroids, obtained by fitting a PRF model to the observed flux, may additionally be computed for some or all targets in the pipeline, though in practice they were computed only for PPA targets. The PRF-based centroiding algorithm determines the photocenter coordinates by solving for the PRF translation and scaling which best fit the cosmic ray-cleaned and background-subtracted flux in the respective target apertures. Flux-weighted centroids are computed in an aperture that includes the optimal aperture plus a single halo ring for each target. PRF-based centroids are computed in an aperture that includes all of the available pixels for each target. In both cases, uncertainties are propagated to the computed centroid row and column coordinates based on the Jacobian for each centroid computation.



**Figure 6.5** Movement of an 11<sup>th</sup> magnitude target's flux-weighted centroid during Q0. These data are from the same target shown in Figure 6.7. The plot on the left shows the path traced on the CCD by the migrating centroid, while the right hand plot shows the row and column centroid time series. The visible data gap near cadence 280 is due to a loss-of-fine-point anomaly accompanying a momentum dump. Transient features in the centroid time series are mainly due to focus changes caused by thermal fluctuations in the photometer, although the brief drops in centroid column lasting  $\sim$ 20 cadences ( $\sim$ 10 hours) every  $\sim$ 90 cadences ( $\sim$ 2 days) is due to an eclipsing binary falling on a fine guidance sensor (FGS) that was selected as a guide star at the time (see Christiansen et al., 2013, Section 4.5). Every time an eclipse happened it biased the FGS centroid measurements, pulling *Kepler* off point by a few millipixels.

Every centroid computed in PA is validated against the bounding box of the associated centroid aperture, and each is gapped in the event that it does not fall within the bounding box for any reason. Gaps are also set for PRF-based

centroids that cannot be successfully computed for any target and cadence due to failure of the iterative fitting algorithm.

### 6.3.4 Motion Polynomial Fitting

In this section we describe in detail the process by which the motion polynomials introduced in Section 6.2 are fitted to the set of PPA target centroids in each unit of work. Even though PPA targets are selected for their likelihood of producing precise and consistent centroids, some of them are unsuitable for this purpose. If the centroid for a given target is determined at run time in PA to be unsuitable for any cadence then the centroids for that target are excluded from the motion polynomial fits for all cadences. Targets exhibiting flux variations greater than 1% are the first major group of unsuitable PPA targets. The centroids for variable targets move with the changing flux level (depending on the degree to which the aperture is crowded and the background has been removed) and will degrade the predictive capability of the motion polynomials if not excluded. Targets for which the reported magnitude appears to be incorrect (based on anomalous uncertainties in their centroids due to contamination from nearby saturated and bleeding bright stars) are the second major group of unsuitable PPA targets. If a target is brighter than its *Kepler* Input Catalog (KIC) magnitude suggests, it may be saturated or affected by a nearby saturated star and therefore completely unsuitable for precision centroiding. If the target is dimmer than its KIC magnitude suggests then the quality of its centroids is also compromised due to lower than expected SNR.

The orders for the 2-D polynomials that are fit to the row and column centroids of the PPA targets for each LC are determined in the pipeline with the AIC in a manner similar to the order selection for the background polynomials. The difference in this case is that the model orders are selected independently for the fits to the centroid rows and columns. The orders used for the polynomial fits to the centroid rows and columns for all cadences are determined by the mode of the orders (up to a parameter-specified maxima) that minimize the AIC for all target rows and columns, respectively, over a decimated set of cadences in the unit of work (in practice, pipeline configurations specified a decimation factor of 8). For a row or column motion polynomial fit of order  $K$ , the number of polynomial coefficients for each cadence is  $(K + 1)(K + 2)/2$ .

As with the background polynomial fitting process, it is necessary to perform the motion polynomial fitting in a robust fashion to prevent centroid outliers from perturbing the least squares solution. The motion polynomial fits are therefore performed iteratively to identify and deemphasize any centroid outliers in row and/or column. Centroid outliers can result, for example, from errors in the target sky coordinates, crowding in the target apertures, bleeding from saturated targets in the vicinity or poor correction of the local background.

For the set  $T$  of PPA targets that have not been excluded from the motion polynomial fit as described above, let  $\mu_r$  and  $\sigma_r$  designate the centroid row coor-

dinates and uncertainties for any given cadence such that  $\mu_{r,t}$  and  $\sigma_{r,t}$  represent the value and uncertainty of the row centroid for the  $t^{th}$  target. Similarly, let  $\mu_c$  and  $\sigma_c$  designate the centroid column coordinates and uncertainties for any given cadence such that  $\mu_{c,t}$  and  $\sigma_{c,t}$  represent the value and uncertainty of the column centroid for the  $t^{th}$  target. Furthermore, let the celestial RA and Dec coordinates of the PPA target  $t$  be designated by  $\alpha_t$  and  $\delta_t$ , respectively. The row motion polynomial  $R$  is then determined for order  $K_R$  by minimizing the weighted  $\chi^2$  defined in Equation 6.9 for the given cadence:

$$\chi_r^2 = \sum_{t \in T} \left( \frac{\mu_{r,t} - \sum_{k=0}^{K_R} \sum_{l=0}^k R_q \alpha_t^{k-l} \delta_t^l}{\sigma_{r,t}} \right)^2. \quad (6.9)$$

Likewise, the column motion polynomial  $C$  is determined for order  $K_C$  by minimizing the weighted  $\chi^2$  defined in Equation 6.10 for the given cadence:

$$\chi_c^2 = \sum_{t \in T} \left( \frac{\mu_{c,t} - \sum_{k=0}^{K_C} \sum_{l=0}^k C_q \alpha_t^{k-l} \delta_t^l}{\sigma_{c,t}} \right)^2. \quad (6.10)$$

The motion polynomial coefficient index  $q$  in both Equation 6.9 and Equation 6.10 is defined in Equation 6.6. Once the weighted least squares problems in Equation 6.9 and Equation 6.10 have been solved for a given cadence, the row and column motion polynomials may then be evaluated for any arbitrary celestial coordinates  $\alpha$  and  $\delta$  (on the associated module output) to estimate the centroid location for a target at the given celestial position, as detailed in Appendix 6-A.

The robust fitting algorithm is performed per cadence and successfully deemphasizes centroid outliers on a per cadence basis, but it does not necessarily consistently weight the centroids from cadence to cadence. The mix of centroids that are effectively fit can therefore potentially vary with time, resulting in discontinuities and chatter in the temporal sequences of the respective motion polynomial coefficients. This could present a problem later in the pipeline where the motion polynomials are utilized for systematic error correction, attitude reconstruction and computation of instrument metrics. The following are enhancements to the robust fitting routine intended to improve both stability and fidelity of the fit coefficients.

To mitigate motion polynomial chatter, the set of targets participating in the fit is determined through an iterative process. Initially the centroids from the set of PPA targets are fit robustly to a 2-D polynomial. Then those targets that have been assigned robust weights (in row or column) below a threshold are removed from the set of targets being fit and the centroids are re-fit. The threshold is configurable, and was set to 100 for the SOC 9.3 processing. This fitting and culling process continues until the set of targets being fit remains unchanged, an iteration limit is reached, or the size of the set of targets being fit drops below a threshold. We also require that the same set of targets be used over all cadences (although the robust weights assigned to any particular target change from cadence to cadence).

Large biases in the centroid data can cause problems with the robust fit as well. Centroids with very large biases may be overly deemphasized in the robust fit, reducing the overall effectiveness of robust weighting. In order to mitigate the effects of any large biases, the residual biases for each centroid time series are first estimated from the fit residuals as a low order polynomial. The model bias is then removed from the original data and the polynomial fit is repeated a second time, providing a more accurate robust fit.

### 6.3.5 Optimal Aperture Selection

PA receives as input a photometric aperture produced by the TAD module (see Chapter 3), shown in Figure 6.1. Since TAD computes optimal apertures without access to the observed pixels, PA attempts to use pixel data to improve the apertures if it can. It constructs a set of candidate apertures, evaluates them along with the original TAD aperture, and selects the best one for use in photometry. Construction and selection of apertures in PA is based on maximizing SNR and minimizing CDPP, as well as on a number of heuristics that render the problem computationally tractable. Construction of candidate apertures begins by fitting a model of image formation to the pixel data within the target mask, enabling isolation of the flux contribution from the target star.

**6.3.5.1 Image Formation Model** An accurate model of image formation is important for SNR estimation because it enables the decomposition of background-subtracted pixel flux into contributions from each point source in the instrument’s field of view. Our model consists of a scene description comprising the set of known sources  $S = \{s_1, s_2, \dots, s_K\}$  with celestial coordinates  $\{\alpha_k, \delta_k\}$  and specified magnitudes, a model of image motion on the focal plane, and a model of the *Kepler* Pixel Response Function (PRF), as illustrated in Figure 6.6. In brief, the PRF describes the expected response of a pixel  $i$  to a point source at an infinite distance whose photocenter falls on the detector at real-valued row and column coordinates  $\{r, c\}$ . In its normalized form, as used here, it describes how the total response due to a point source is distributed among pixels. The PRF accounts for the effects of the optical PSF, the CCD detector responsivity function, spacecraft pointing jitter, and other systematic effects. A static model of the *Kepler* PRF was constructed from 121 dithered LC data sets acquired during commissioning of the *Kepler* spacecraft (Bryson et al., 2010a).

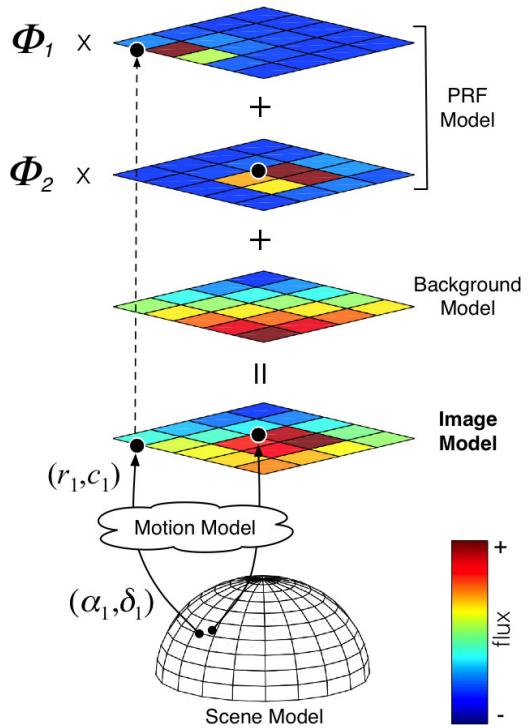
After subtracting the estimated background flux  $\hat{\phi}_b$  and cosmic ray flux  $\hat{\phi}_{cr}$  from the calibrated pixel flux  $\phi$ , we model the remaining flux,  $\phi_s$ , at each pixel  $i$  during cadence  $n$  as a linear combination of PRF values for each source plus a constant offset:

$$\phi_s(i, n) = \beta(n) + \sum_{k=1}^K \Phi_k(n) \widehat{PRF}(i, r_k(n), c_k(n)). \quad (6.11)$$

The total flux  $\Phi_k$  due to source  $s_k$  is distributed among pixels according to the normalized PRF, denoted by  $\widehat{PRF}$ .  $\widehat{PRF}$  is obtained by evaluating the PRF over an 11-by-11 pixel grid centered at  $\{r, c\}$  and dividing pixel  $i$  by the sum of evaluated pixels. The real-valued CCD coordinates  $r_k(n)$  and  $c_k(n)$  are estimates of the mean photocenter of source  $s_k$  during cadence  $n$  and are obtained from motion polynomials as detailed in Appendix 6-A. The constant term  $\beta$  is included to account for any local bias in the background flux estimates for the set of pixels being modeled.

An independent model is fitted to each set of pixels comprising a target mask. The scene model for a given mask is constructed from entries in the KIC and is initialized by ranking sources in the vicinity of the target by their maximum expected SNR contribution to any pixel in the mask. Sources whose maximum SNR values fall below a threshold of 100 are discarded and up to  $K_{max} = 30$  sources from the remaining set are admitted to the scene model in order of rank. Since certain regions of the FOV may be densely populated with dim stars, limiting the scene model in this way prevents wasting time and computing power on sources that do not make significant flux contributions.

We fit model parameters  $\Phi_k(n)$ ,  $\beta_k(n)$ ,  $\alpha_k$  and  $\delta_k$  to the observed data by minimizing the weighted sum of chi-square residuals over all pixels and cadences in the model. Fine-tuning the source positions  $\{\alpha_k, \delta_k\}$  in the fit compensates for both catalog errors and local biases in the motion polynomials. Position fitting is enabled only for sources with centroids inside the target mask that contribute sufficient flux, defined by a maximum pixel SNR threshold of 500. The minimization is constrained such that source flux is non-negative and perturbations  $\Delta_k$  to catalog positions are smaller than  $\Delta_{max} = 1.5$  pixels. Note that, while fitting  $\Phi_k$  and  $\beta$  is a linear problem, fitting source positions  $\alpha_k$  and  $\delta_k$  is inherently nonlinear.



**Figure 6.6** A single-cadence model of an actual *Kepler* target pixel mask. The image model is constructed from a background model, a normalized PRF model evaluated for each source, and estimates  $\Phi_k$  of the total flux contributed by each source. The celestial coordinates  $(\alpha_k, \delta_k)$  of each source are mapped to CCD coordinates  $(r_k, c_k)$  by a polynomial model of image motion. Note that in this example the centroid of source  $s_1$  lies slightly outside the mask, but it still contributes significant flux to pixels inside the mask. Also note that the component images in the figure have been scaled to aid visualization. From Figure 1 in Smith et al. (2016).

We address the two components of the fit separately in an iterative scheme. We use a Levenberg-Marquardt algorithm to optimize the perturbation of source positions, while a non-negative least squares method (Kim et al., 2013) is used to fit the flux model of Equation 6.11 for a given perturbation:

$$\begin{aligned} & \text{minimize} && (w(\Delta) + 1) \sum_{i=1}^I \sum_{n=1}^N \chi^2(i, n; \Phi(n), \beta(n), \Delta) \\ & \text{subject to} && \Phi_k(n) \geq 0, \Delta_k \leq \Delta_{max}, \end{aligned} \quad (6.12)$$

where

$$\chi(i, n) = \frac{\phi_s(i, n) - \hat{\phi}_s(i, n)}{\sigma_s(i, n)}, \quad (6.13)$$

and  $\sigma_s(i, n)$  denotes the uncertainty associated with  $\phi_s(i, n)$ . The heuristic weighting function  $w(\Delta)$  in Equation 6.12 prevents source positions from converging on the same peak in the data by increasing as sources move toward one another from their catalog positions.

The resulting image model provides the required estimate of flux due to the target star in isolation, which is given by:

$$\hat{\phi}_t(i, n) = \Phi_t(n) \widehat{PRF}(i, r_t(n), c_t(n)). \quad (6.14)$$

It also enables easy calculation of flux fraction and crowding metrics (see Sub-section 6.3.6 for detailed descriptions of these quantities) for each cadence.

**6.3.5.2 Aperture Optimization** For pixel  $i$  and cadence  $n$ , the modeled target flux of Equation 6.14 provides the numerator for an estimate of SNR. The denominator, or noise term, is the sum of estimated shot noise, read noise, and quantization noise (Smith et al., 2016; Jenkins et al., 2000; Howell, 1989). The contiguous aperture that maximizes SNR is then identified by way of a region-growing algorithm. The pixel containing the modeled target centroid  $\{r_t(n), c_t(n)\}$  provides the seed, which we designate  $A_1(n)$ . A second region,  $A_2(n)$ , is constructed by identifying the 4-connected neighbor of  $A_1(n)$  having maximum SNR. In general, the  $p$ -pixel region  $A_p(n)$  is obtained by finding the 4-connected neighbor of  $A_{p-1}(n)$  having maximum SNR. This procedure is repeated until all pixels in the  $P$ -pixel target mask belong to at least one region and we have a set of cadence-specific candidate apertures  $A_1(n), A_2(n), \dots, A_P(n)$ . Note that these apertures are not necessarily symmetric and may contain holes. The SNR of  $A_p(n)$  is given by:

$$\text{SNR}_p(n) = \frac{\sum_{i \in A_p(n)} \hat{\phi}_t(i, n)}{\sqrt{p (\nu_{\text{read}}(n)^2 + \nu_{\text{quant}}(n)^2) + \sum_{i \in A_p(n)} \phi_s(i, n)}}, \quad (6.15)$$

where  $\hat{\phi}_t(i)$  denotes the modeled target flux at pixel  $i$  and  $\phi_s(i)$  denotes the corrected pixel value defined in Equation 6.8, which includes flux contributions

from all background stars (the shot noise term is derived from these values). The other two terms,  $\nu_{\text{read}}$  and  $\nu_{\text{quant}}$ , are the read noise and quantization noise, respectively.

For each cadence we select the candidate aperture that maximizes SNR. We now have an optimal (in the SNR sense) photometric aperture for each cadence, denoted by  $A_{\text{opt}}(n)$ , but the SAP method employed by the *Kepler* pipeline requires a single static (i.e., time-invariant) aperture to be used for photometry on all cadences. There are many ways we can construct such an aperture. Our approach is once again to construct several candidates and then select the best.

Two of our static aperture candidates are derived directly from the set of per-cadence optimal apertures,  $A_{\text{opt}}(n)$ . This is done by counting how many of the  $A_{\text{opt}}(n)$  a given pixel belongs to and adding the pixel to the static aperture if its frequency of inclusion is above a threshold percentage. We do this for thresholds of 5% and 50%. Since the candidate resulting from the 5% threshold can be seen as the union of 95% of the  $A_{\text{opt}}(n)$ , we refer to it as the *95% union aperture* and denote it by  $A_{\text{union}}$ . The 50% threshold produces an aperture that can be considered the median of the  $A_{\text{opt}}(n)$  apertures and is denoted by  $A_{\text{median}}$ . For targets with small centroid motion, the median aperture is optimal since it finds the “core” of pixels containing the signal and is robust to outliers in the centroid position. When there is large motion, however, a significant fraction of the target flux is distributed to outlying pixels not included in  $A_{\text{median}}$ . In such situations the larger  $A_{\text{union}}$  is often preferred.

A third candidate aperture is constructed by minimizing estimated CDPP. Since it is not computationally feasible to perform the minimization over all possible apertures, we do so for a limited subset. Letting  $m(i)$  denote the number of the  $A_{\text{opt}}(n)$  to which pixel  $i$  belongs, we assign pixel  $i$  a rank equal to  $m(i)$ . Starting with the central pixel we construct a subset of apertures of increasing size by adding pixels in order of their rank (highest to lowest). By doing so we reduce the number of possible apertures from  $2^P - 1$  (all possible subsets of pixels minus the empty set) to  $P$ . We can now afford to compute the much more costly estimate of CDPP for each of the  $P$  candidate apertures to find the aperture with the lowest CDPP.

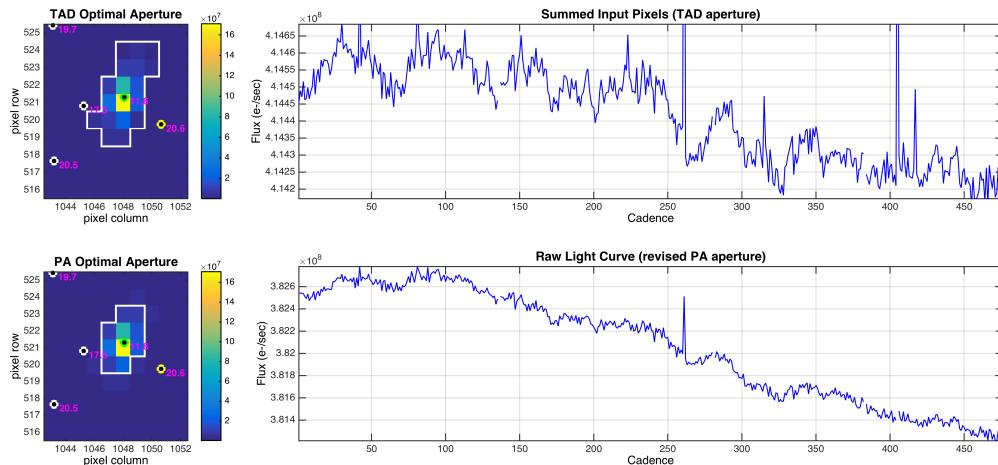
At this point we have four candidate apertures from which we must select one:

1.  $A_{\text{TAD}}$ , the purely model-derived aperture computed in TAD.
2.  $A_{\text{median}}$ , the SNR-optimized median aperture.
3.  $A_{\text{union}}$ , the SNR-optimized 95% union aperture.
4.  $A_{\text{CDPP}}$ , the CDPP-optimized aperture.

Each of these apertures can yield better performance for a subset of *Kepler* targets. Since transit signal detection is the primary goal of the *Kepler* mission, we principally select the optimal aperture based on CDPP. It may seem that the CDPP-optimized aperture should always be the best choice. But even if we were

to optimize CDPP over all possible apertures, which is not computationally feasible, there would still be cases in which a purely CDPP-based selection would lead to poor results. For targets with bright background objects in close proximity, a CDPP-based optimization could engulf the background object, depending on the attributes of its light curve. We therefore select from the four optimal aperture candidates using a strategy that combines CDPP and SNR by way of a tuned logistical regressive heuristic model (Smith et al., 2016) detailed in Chapter 7.

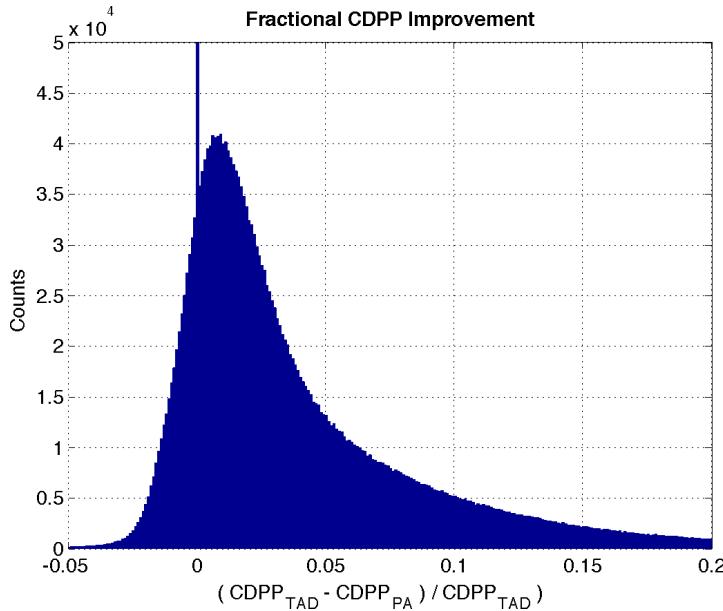
Figure 6.7 shows an example of an optimal aperture selected in the manner described here alongside the TAD-computed optimal aperture passed as input to PA. The histogram shown in Figure 6.8 indicates that revised apertures reduce the estimated CDPP for the vast majority of targets and quarters, with 77% showing an improvement, 13% showing some degradation, and 10% showing no change. In 14% of cases CDPP is reduced by 10% or more.



**Figure 6.7** Example of Q0 pixel/target flux inputs and outputs for an 11<sup>th</sup> magnitude target (KIC 10645746) near the edge of the focal plane (module output 6.3). The top half of the figure shows PA inputs, including the pixels comprising the target mask (top left) with the TAD optimal aperture bounded in white and the sum of calibrated flux time series (top right). The bottom half shows PA outputs, including the revised optimal aperture along with the raw light curve delivered to the MAST archive. The aperture plots on the left side of the figure show the median pixel flux values in the target mask over 476 cadences. KIC objects are represented by circular markers with the target star in green and background objects in white or yellow. Objects are shown at their median (over all cadences) CCD coordinates as predicted by motion polynomials. The *Kepler* magnitude of each object is shown in magenta to the right of its marker. Features to note include the exclusion from the revised aperture of pixels near background stars and the removal of most positive-going spikes due to cosmic rays and Argabrightenings in the output light curve.

### 6.3.6 Simple Aperture Photometry

Storage and bandwidth limitations made it impossible to save and later downlink all of the pixel values acquired on the focal plane array on every cadence. Rather, an aperture was defined for each target that specified the pixels necessary to support pixel-level calibrations, light curve extraction, and computation of the target



**Figure 6.8** Histogram of fractional CDPP improvement when comparing light curves produced from revised PA apertures with those produced from the original TAD apertures. Positive values indicate improved photometric precision. Note that each target is counted separately for each quarter. That is, the sum of the histogram is (number of targets)  $\times$  (number of quarters).

photocenter. Only pixels in the specified target apertures (plus the background and collateral (Quintana et al., 2010) pixels for black level and smear corrections in CAL) were written to the solid state recorder aboard the spacecraft and downlinked for pipeline processing. Great care was taken in generation of the aperture definitions (Bryson et al., 2010b) to ensure that all of the pixels required to support the *Kepler* science mission were captured.

Within each target aperture, the subset of pixels required for photometric extraction of light curves is referred to as the optimal aperture. The size of the optimal aperture for any given target depends on a number of factors, including target magnitude, PRF, noise level, local crowding, image motion, and DVA. The target apertures were redefined for each observing season as the targets moved from one CCD to another with each quarterly roll of the photometer. An *a priori* optimal aperture is estimated for each target by the TAD component (Figure 6.1) in order to identify the pixels that need to be collected and stored onboard the spacecraft for later downlink. This estimate is provided to PA as input along with the pixel data and is used as the initial estimate for the photometric aperture. After the motion polynomials are determined in the first run of PA, TAD is called again to furnish updated optimal apertures using the reconstructed image motion across the observation period. As described in Subsection 6.3.5, PA uses the observed pixels to update the estimated optimal aperture via PA-COA and can fall back on the TAD aperture should any problems arise in its analysis. Be-

fore the introduction of PA-COA in release 9.3, the updated TAD apertures were employed to compute SAP flux.

The optimal aperture does not necessarily contain all of the stellar flux for a given target. Along with the raw light curve, PA delivers two additional values at each cadence: the *flux fraction* and *crowding metric*. The flux fraction in the aperture refers to the ratio of target flux contained in the optimal aperture to the total flux of the target. Furthermore, not all flux in the optimal aperture is due to the primary target. The crowding metric refers to the fraction of flux in the optimal aperture that is due to the target. Both flux fraction and crowding are accounted for when the light curves are corrected in PDC (Twicken et al., 2010a).

As stated in Section 6.2, light curves are computed by SAP. Once the calibrated target pixels have had cosmic rays corrected and background removed, the raw flux is obtained per target and cadence by the unweighted summation of pixels in the associated optimal aperture. Letting  $A_{opt}$  denote the set of indices for pixels within the optimal aperture of a specified target, the raw flux  $\Phi$  is computed by SAP from the background removed pixels  $\phi_s$  in the target aperture for a given cadence by:

$$\Phi = \sum_{i \in A_{opt}} \phi_s(i). \quad (6.16)$$

The SAP results from an 11<sup>th</sup> magnitude target star are shown in the lower panels of Figure 6.7. It should be noted that if a data gap exists for any pixel and cadence in the optimal aperture of a given target, then a data gap is set for the raw flux value (and associated uncertainty) for that target and cadence. Setting the gaps in this manner prevents discontinuities from being introduced into the light curves by extracting the raw flux from only a subset of the pixels in the optimal aperture. As this is written, it has not been observed in-flight that data availability is pixel dependent. Rather, it has always been true that either all or none of the pixels are valid for a given cadence.

## 6.4 Summary and Conclusions

We have presented an overview of the architecture and algorithms of the Photometric Analysis component of the *Kepler* SOC Science Data Processing Pipeline. Detailed expositions of the main science algorithms and examples of the primary data products were given. We have provided examples of the main PA data products and have shown a significant improvement in estimated CDPP of light curves due to refinement of photometric apertures.

## Appendix A: Obtaining Centroid and Plate Scale Estimates from Motion Polynomials

For a given cadence, the row centroid estimate is obtained by evaluating the order  $K_R$  row polynomial with coefficients  $R_q$  at the target's celestial coordinates  $(\alpha, \delta)$ ,

$$r = \sum_{k=0}^{K_R} \sum_{l=0}^k R_q \alpha^{k-l} \delta^l, \quad (\text{A.1})$$

where  $q = k(k + 1)/2 + l$ , and likewise for the column centroid estimate:

$$c = \sum_{k=0}^{K_C} \sum_{l=0}^k C_q \alpha^{k-l} \delta^l. \quad (\text{A.2})$$

Plate scale is computed from the partial derivatives of the motion polynomials at each cadence with respect to right ascension and declination by calculating the determinant of the Jacobian matrix:

$$s = \begin{vmatrix} \frac{\partial r}{\partial \alpha} & \frac{\partial r}{\partial \delta} \\ \frac{\partial c}{\partial \alpha} & \frac{\partial c}{\partial \delta} \end{vmatrix}. \quad (\text{A.3})$$

The partial derivatives in Equation A.3 are given by

$$\begin{aligned} \frac{\partial r}{\partial \alpha} &= \sum_{k=0}^{K_R} \sum_{l=0}^k R_q [(k - l) \alpha^{k-l-1}] \delta^l, \\ \frac{\partial r}{\partial \delta} &= \sum_{k=0}^{K_R} \sum_{l=0}^k R_q \alpha^{k-l} [l \delta^{l-1}], \\ \frac{\partial c}{\partial \alpha} &= \sum_{k=0}^{K_C} \sum_{l=0}^k C_q [(k - l) \alpha^{k-l-1}] \delta^l, \text{ and} \\ \frac{\partial c}{\partial \delta} &= \sum_{k=0}^{K_C} \sum_{l=0}^k C_q \alpha^{k-l} [l \delta^{l-1}]. \end{aligned} \quad (\text{A.4})$$

## Bibliography

- Akaike, H., 1974. "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, 19, 716
- Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al., 2010. "The Kepler Pixel Response Function," ApJL, 713, L97

- Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010b. "Selecting Pixels for Kepler Downlink," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401D
- Burg, J. P. 1967. "Maximum Entropy Spectral Analysis," in Proc. 37th Meeting Soc. Exploration Geophys.
- Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al., 2012. "The Derivation, Properties, and Value of Kepler's Combined Differential Photometric Precision," PASP, 124, 1279
- Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2013. Kepler Data Characteristics Handbook (KSCI-19040-004) (Moffett Field, CA: NASA Ames Research Center)
- Clarke, B. D., Allen, C., Bryson, S. T., et al. 2010. "A Framework for Propagation of Uncertainties in the Kepler Data Analysis Pipeline," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774020
- Howell, S. B., 1989. "Two-Dimensional Aperture Photometry - Signal-to-Noise Ratio of Point-Source Observations and Optimal Data-Extraction Techniques," PASP, 101, 616
- Jenkins, J. M., Caldwell, D. A., & Gilliland, R. L. 2004. KeplerAlgorithm Theoretical Basis Document: KSOC-21008 (Moffett Field, CA: NASA Ames Research Center)
- Jenkins, J. M., Witteborn, F., Koch, D. G., et al. 2000. "Processing CCD Images to Detect Transits of Earth-Sized Planets: Maximizing Sensitivity while Achieving Reasonable Downlink Requirements," in Proc. SPIE, Vol. 4013, UV, Optical, and IR Space Telescopes and Instruments, ed. J. B. Breckinridge & P. Jakobsen, 520–531
- Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al., 2010. "Initial Characteristics of Kepler Long Cadence Data for Detecting Transiting Planets," ApJL, 713, L120
- , 2010. "Overview of the Kepler Science Processing Pipeline," ApJL, 713, L87
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010c. "Transiting Planet Search in the Kepler Pipeline," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D
- Kim, D., Sra, S., & Dhillon, I. S., 2013. "A Non-Monotonic Method for Large-scale Non-negative Least Squares, journal = Optimization Methods Software," 28, 1012
- Klaus, T. C., McCauliff, S., Cote, M. T., et al. 2010a. "Kepler Science Operations Center Pipeline Framework," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774017

- Klaus, T. C., Cote, M. T., McCauliff, S., et al. 2010b. “The Kepler Science Operations Center Pipeline Framework Extensions,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774018
- Li, J., Allen, C., Bryson, S. T., et al. 2010. “Photometer Performance Assessment in Kepler Science Data Processing,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401T
- Papoulis, A. 1986. Probability, Random Variables, and Stochastic Processes (McGraw-Hill)
- Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010. “Pixel-Level Calibration in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401X
- Smith, J. C., Morris, R. L., Jenkins, J. M., et al., 2016. “Finding Optimal Apertures in Kepler Data,” PASP, 128, 124501
- Stumpe, M. C., Smith, J. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning I – Architecture and Algorithms for Error Correction in Kepler Light Curves,” PASP, 124, 985
- Tenenbaum, P., & Jenkins, J. M. 2010. “Focal Plane Geometry Characterization of the Kepler Mission,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401C
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Twicken, J. D., Chandrasekaran, H., Jenkins, J. M., et al. 2010a. “Presearch Data Conditioning in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401U
- Twicken, J. D., Clarke, B. D., Bryson, S. T., et al. 2010b. “Photometric Analysis in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774023
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)
- Witteborn, F. C., Van Cleve, J., Borucki, W., Argabright, V., & Hascall, P. 2011. “DEBRIS Sightings in the Kepler Field,” in Proc. SPIE, Vol. 8151, Techniques and Instrumentation for Detection of Exoplanets V, 815117
- Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010. “Data Validation in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, 42W

## CHAPTER 7

---

# FINDING OPTIMAL APERTURES IN KEPLER DATA

JEFFREY C. SMITH<sup>1</sup>, ROBERT L. MORRIS<sup>1</sup>, JON M. JENKINS<sup>2</sup>, STEPHEN T. BRYSON<sup>2</sup>, DOUGLAS A. CALDWELL<sup>1</sup>, AND FORREST R. GIROUARD<sup>3</sup>

<sup>1</sup> The SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035, <sup>2</sup> NASA Ames Research Center, Moffett Field, CA 94035, <sup>3</sup>Logyx LLC/NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** With the loss of two spacecraft reaction wheels precluding further data collection in the *Kepler* primary mission, even greater pressure is placed on the Science Data Processing Pipeline to eke out every last transit signal in the data. To that end, we have developed a new method to optimize the *Kepler* Simple Aperture Photometry (SAP) photometric apertures for both planet detection and minimization of systematic effects. The approach uses a per cadence modeling of the raw pixel data and then performs an aperture optimization based on Signal-to-Noise ratio (SNR) and the *Kepler* Combined Differential Photometric Precision (CDPP), which is a measure of the noise over the duration of a reference transit signal. We have found the new apertures to be superior to the previous *Kepler* apertures. We can now also find a per cadence Flux Fraction in Aperture and Crowding Metric. The new approach has also been proven robust at finding apertures in K2 data that help mitigate the larger motion-induced systematics in the photometry. The method allows us to identify errors in the *Kepler* and K2 input catalogs. This chapter draws on an updated version of Smith et al. (2016).

**Keywords:** Stars; Extrasolar Planets; Data Analysis and Techniques

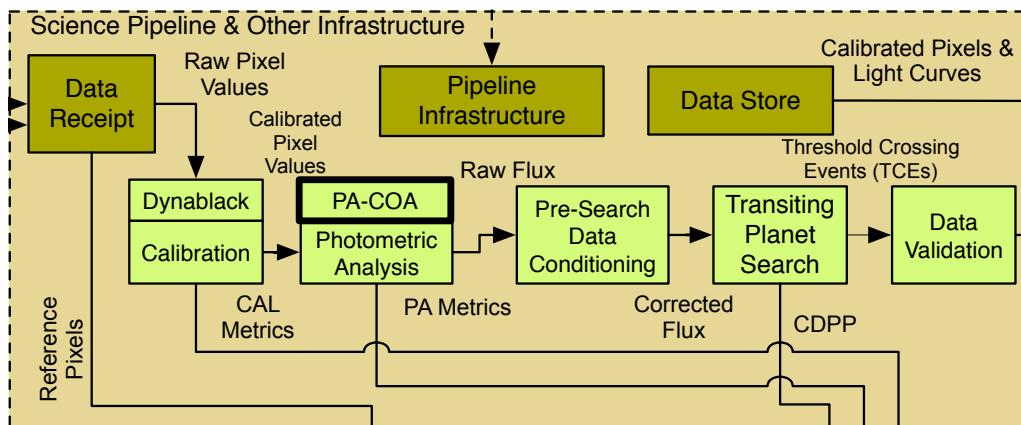
### 7.1 Introduction

To achieve Kepler's primary goal of discovering potentially habitable Earth-size planets transiting Sun-like stars in its 116 square degree field of view (FOV), obtaining photometric precision is vital. The task of determining which pixels are used to formulate the photometric measurement for the *Kepler* target stars is a crucial component of this effort. Prior to the final software release, SOC 9.3, the photometric apertures were determined by a component called Create Optimal Apertures (COA) as part of the Target Management function of the SOC pipeline within the software module Target and Aperture Definitions (TAD; see Chapter 3). COA performed two distinct jobs:

1. It identified the pixels that needed to be captured and stored onboard *Kepler*'s solid state recorder (SSR) based on a predicted pointing profile and models for the CCDs, point spread function (PSF), and sky.

- It identified the pixels to be used for extracting photometry from the pixels that were stored and downlinked to the ground, based on the actual pointing behavior.

The TAD module relied on the accuracy and validity of the *Kepler* Input Catalog (KIC) and its updates to formulate valid apertures. In practice, a subset of stars had catalog errors in magnitude and celestial coordinates as well as right ascension and declination ( $\alpha, \delta$ ), resulting in sub-optimal apertures and degraded flux time series. This motivated the development of a new module called PA-COA in SOC 9.3 that allowed the actual pixel data to be used in formulating the photometric apertures, in addition to the catalog, focal plane models, and reconstructed pointing histories. Figure 7.1 shows where PA-COA fits in the context of the SOC Pipeline.



**Figure 7.1** PA-COA in the context of the architecture of the SOC. PA-COA determines the apertures used in the Photometric Analysis (PA) module to produce simple aperture photometry from the calibrated pixels flowing into PA from the calibration module (CAL).

## 7.2 Finding Optimal Apertures

Due to bandwidth considerations, data from all of the 95 million pixels in the *Kepler* focal plane cannot be stored for downlink for every cadence. Therefore specific objects (mostly stars) in the *Kepler* FOV are identified as targets.<sup>1</sup> Groups of pixels around these objects are called masks and are extracted for storage and eventual transmission. Within each mask is a smaller collection of pixels called an “aperture” that is optimized for generating photometric light curves. For some targets, such as galaxies, the required apertures are explicitly specified and not optimized in the pipeline. For planetary transit targets the pixels in the aperture are selected to maximize the SNR and photometric precision for the target by

<sup>1</sup>We will use the term “source” to reference any individual point source in the FOV be it a star, galaxy or otherwise. The targets are almost all stars but a small number of targets are clusters, galaxies or AGNs.

TAD. The resulting collection of pixels is called an “optimal aperture”. For the *Kepler* Pipeline a single fixed optimal aperture is found for each data quarter of approximately 90 days. The maximum motion for *Kepler* data is  $\tilde{0.6}$  pixels over a 90-day quarter, with an average of 0.4 pixels, plus minor pointing drift. Although using moving adaptive apertures is unavoidable for many ground-based photometric surveys, where the image motion can be several pixels and upwards over a night of observations, it is not necessary with the pointing precision similar to that of *Kepler*.

Two principle methods are used for finding the optimal apertures. One is model-driven, the other data-driven. The first method (Method #1) generates two Pixel Response Function (PRF) model-derived synthetic Full Frame Images (FFIs), one with and one without the target star (the PRF is discussed in Section 7.3 and in Chapter 3). These FFIs are used to estimate the SNR of each pixel, allowing the selection of the pixel set that maximizes the SNR of the total flux of each target. The second method (Method #2) uses a PRF model image fit to the acquired pixel data to find the target flux contribution to the scene. The method then uses the pure pixel scene data to estimate the noise. This second method has a further step where the aperture is optimized with Combined Differential Photometric Precision (CDPP, discussed in Subsection 7.4.2). The quality of the resultant aperture and light curve is compared between the two methods and the better of the two is chosen for each target. Method #1 has been used for the *Kepler* Mission since before launch and is how apertures are selected in flight. A brief description of this method is presented in Section 7.3 for reference. Method #2 is novel to the SOC 9.3 software release.

### 7.3 Method #1: Using the Synthetic FFI and a Pure PRF Image Model

In this method the computation of optimal apertures is based on the generation of two synthetic FFIs for each CCD channel without the use of the acquired pixel data, one FFI with all stars and a second FFI with the target star removed (though still including the effects of the target star, such as smear – see Subsection 3.2.1 and Subsection 3.2.2). These images are then used to compare the signal from the target star with the noise from the target star, stellar background, and the instrument thereby optimizing the aperture for SNR. By considering all the pixels in these two FFIs in a region around a target, we can determine the pixel set whose sum maximizes the SNR for that target.

The undersampled *Kepler* Point Spread Function (PSF) combined with the intra-pixel variability causes the PRF to be very sensitive to small motions of a point source’s centroid<sup>2</sup>. These motions include spacecraft pointing jitter and differential velocity aberration and can be a significant source of signal noise and uncertainty. Therefore it is important to characterize each pixel’s response

<sup>2</sup>The PRF is approximately the convolution of the PSF with the responsivity of a pixel over its spatial extent (see Bryson et al., 2010a).

to image motion on a sub-pixel scale. We call the function describing a pixel’s response to the location of a point source the Pixel Response Function (PRF) (Bryson et al., 2010a).

The PRF models were determined during commissioning using 121 long-cadence data sets acquired at a 15-minute cadence (Bryson et al., 2010a). They account for components of the mean motion due to spacecraft pointing jitter and other effects that occur on a single long-cadence time scale. The PRFs depend sensitively on the sub-pixel location of the point sources. We therefore approximate this dependence by characterizing the PRF as a piece-wise 2-D polynomial on a  $6 \times 6$  sub-pixel grid, so each pixel has 36 2-D polynomials representing the PRF waveform in that pixel, creating a super-resolution representation of the PRF on a 0.17 pixel grid.<sup>3</sup> For each sub-pixel the polynomial was computed as a robust  $\chi^2$  fit to the normalized pixel fluxes for all pixels of stars in that CCD readout region falling in that sub-pixel using the commissioning data, as discussed in (Bryson et al., 2010a). Five PRFs were determined for each CCD readout channel in order to capture variation in the focus across each CCD, each representing the PRF in a specific region of the CCD channel, with one for the center, and one for each of the corners. To evaluate the PRF for a particular star, the three of the five canonical PRFs for that readout channel that contain the CCD region in which the stellar image falls are interpolated to obtain a PRF model specific to that star’s location. We use the commissioning PRFs, the target catalog and estimates of the image motion ( $\Delta x, \Delta y$ ) over time due to pointing and differential velocity aberration (DVA) to compute a synthetic FFI, which is then used to estimate the set of pixels that provide an optimal signal for a target.

The sub-pixel grid accounts for intra-pixel variability. For each sub-pixel location the PRF is defined on a domain of up to  $15 \times 15$  pixels centered on the point source centroid. The PRF model thus found is represented by interpolating a set of polynomial coefficients expressing the brightness of each pixel as a function of offsets of the light falling on that pixel by convolving the flux centered on each of that pixel’s sub-pixel position with the coefficients of the pixel response function for that sub-pixel position, and summing over the sub-pixel positions. Motion over longer time scales is explicitly accounted for by making the offsets  $\{\Delta x(t), \Delta y(t)\}$  time dependent.

The computation of a synthetic FFI uses several inputs (Bryson et al., 2010a):

- ***Kepler* Input catalog (KIC) (Brown et al., 2011)**, providing J2000 right ascension, declination and magnitude in the *Kepler* bandpass of stellar targets in the *Kepler* field.
- **Pixel Response Function (PRF) model**, an observation-based super-resolution model of expected pixel values in response to point sources. The PRF model includes intra-pixel variability.

<sup>3</sup>Note that the PRF models located on the MAST are resampled versions of the piecewise-polynomial PRFs used internally by the SOC, and are documented in the KAM (Thompson et al., 2016).

- **Focal Plane Geometry (FPG) and pointing model**, which includes measurements of the locations of the CCDs in the *Kepler* focal plane, models of the *Kepler* optics and of differential velocity aberration (DVA).
- **Saturation model**, which includes information about the well depth of each output channel.
- **Zodiacal light model**, represented as a mesh of magnitude values on the sky.
- **Read noise model**, observed values of read noise for each output channel.
- **Charge Transfer Efficiency (CTE) model**, which describes how much flux is lost with each parallel or serial charge transfer during readout.

In essence, for each source in the KIC that falls on a channel, the source's pixel position on the channel is computed from the KIC, using the FPG and pointing model and a copy of the appropriate PRF scaled by the brightness of that source and smeared by DVA motion. This is all added to the FFI in the appropriate pixel location. Due to the large number of sources in the KIC and the resolution required to estimate the DVA motion, the direct computation of the FFI would be impractically slow. To overcome this difficulty, both the PRF and the brightness of each pixel in the FFI are represented by two-dimensional polynomials, and most of the computations are performed in terms of the polynomial coefficients. The contribution from each individual source is included in the model by adding the coefficients of the PRF's polynomial representation to the coefficients of the FFI pixel's polynomial representation. Once this process is complete the FFI is generated by evaluating the FFI's polynomial for each pixel. The synthetic FFI is completed by adding simulated spill-over of saturated pixels, effects of charge transfer, smear due to shutterless operation and zodiacal light.

The above computation of a synthetic FFI uses all known sources that fall on a channel. A subset of these sources are identified as target stars and the same procedure is used to generate synthetic images of each target's contribution in isolation. The isolated target images are then subtracted from the all-inclusive FFI to produce an FFI with the target flux removed. When removing the flux from a target star to compute the second FFI, care is taken to retain the smear signal from all other sources. A noise model is then used to estimate the noise in each pixel. This noise model includes shot noise of the target, background signal, smear and zodiacal light as well as read and quantization noise. The SNR for each pixel is computed using

$$\text{SNR}_{\text{pixel}} = \frac{f_{\text{target}}}{\sqrt{f_{\text{target}} + f_{\text{back}} + \nu_{\text{read}}^2 + \nu_{\text{quant}}^2}}, \quad (7.1)$$

where  $\nu_{\text{read}}$  is the channel's read noise based on the read noise model and  $\nu_{\text{quant}}$  is the quantization noise, given by

$$\nu_{\text{quant}} = \sqrt{\frac{n_C}{12}} \left( \frac{w}{2^{n_b-1}} \right)^2, \quad (7.2)$$

where  $n_C$  is the number of cadences in a co-added observation ( $n_C = 270$  for *Kepler* Long Cadence),  $w$  is the well depth and  $n_b$  is the number of bits in the analog-to-digital converter (14). The other two terms in Equation 7.1,  $f_{\text{target}}$  and  $f_{\text{back}}$ , are the target and background flux values from the FFI and account for the Poisson shot noise (which scales as the square root of the flux).

Given a collection of pixels, the SNR of the collection is given by Equation 7.1 summed over the pixels in the collection. Optimal pixel selection begins by including the pixel with the highest SNR. The next pixel to be added is the pixel that results in the greatest increase in SNR of the collection. Initially the collection SNR will increase as pixels are added. After the bright pixels in the target have been added, dim pixels dominated by noise cause the SNR to decrease. The pixel collection with the highest SNR defines the optimal aperture.

This computation of the optimal aperture is based on a method originally implemented in the *Kepler* End-to-End Model (ETEM), described in Jenkins et al. (2004) with modifications described in Bryson (2008) and Bryson et al. (2010b) (see also Chapter 3). It is a model-driven approach and does not use the actual CCD pixel data in the computation. It relies on an accurate PRF, pointing knowledge, target catalog and representation of noise sources. Herein lies its main deficiency: errors in the PRF model or target catalog position and magnitude directly result in aperture errors. Incomplete accounting for background objects by the catalogs used, focus errors, stellar variability and saturation also impair the method. While this method is fast and reliably selects an aperture, the above errors lead to compromised photometry on some targets. Using collected pixel data to identify and correct these errors is the motivation for the method described in the next section.

#### 7.4 Method #2: Using a PRF-Based Image Model and the Pixel Scene

The driving factor to improve upon the older aperture finding routine is to allow the data to speak for itself when finding both the background noise and the target signal. The *Kepler* catalog is quite complete to 17th magnitude but dimmer background objects can still contribute to the noise and yet not be modeled in Method #1. There can also be errors of up to  $\pm 0.5$  magnitudes and  $4''$  in position (Brown et al., 2011) for known catalog objects which further contributes to errors in the found optimal apertures.

Method #2 fits an image model to the scene and can thereby update the catalog targets' positions and brightnesses. It then calculates the signal flux for each pixel in the mask based on this model, which is then used as the numerator in

the SNR ratio. Note that, although the image model exploits the existing PRF models, this is not PRF photometry per se, as we are not generating light curves from the PRF model but merely using the modeled image for the numerator in the SNR calculation.

The optimal aperture Method #2 algorithm includes a multi-step process. For each cadence,

1. The PRF model is fit to the pixel scene using the catalog as initial values and then determines the contribution to the pixel flux from each catalog object plus the residual background. A description of the image modeling is given in Subsection 7.4.1 below.
2. The target center pixel is found based on the PRF model fit and this pixel is labeled as the first pixel.
3. The pixel adding order is found by calculating the SNR for each progressively larger aperture size and always choosing the next pixel that increases the SNR the most. This will order the pixels by their contribution to the SNR.
4. The peak in the SNR curve is found using Equation 7.3 below; this is the SNR-maximized optimal aperture for this cadence.

This part of Method #2 is very similar to Method #1 except for the use of real data. So, instead of using Equation 7.1, here we compute the SNR on a per cadence basis using the following equation:

$$\text{SNR}_k = \frac{\sum_i^k f_i}{\sqrt{k (\nu_{\text{read}}^2 + \nu_{\text{quant}}^2) + \sum_i^k y_i}}, \quad (7.3)$$

where  $f$  is the target image fit to the PRF model (as given in Equation 7.4 below),  $y$  is the *actual pixel data* including all background (this is the shot noise term), the other two terms are the non-shot noise (read noise & quantization noise) as computed in Section 7.3 and  $i$  is over each pixel in the aperture of size  $k$ . Note that the calibrated pixel data produced by the calibration (CAL) module are used (i.e., corrected for bias, flat-field, etc.). When ordering the pixels there is no requirement for symmetric apertures and there can be holes in the aperture; however the apertures must be contiguous in the 4-connected sense (i.e., each pixel in the aperture has a solid edge in contact with at least one other pixel in the aperture).

We now have an optimal aperture for each cadence. The *Kepler* Mission pipeline creates photometric light curves using *simple aperture photometry*, which means a single aperture is used for each quarter. There are different ways we can construct the single optimal aperture and we find two apertures directly from this SNR-optimized, per-cadence aperture. The first is a 95% union of each cadence's aperture. That is to say, a pixel is included if greater than 5% of the

cadences include that pixel in its optimal aperture. The other is a 50% union, or median aperture over all cadences. For targets with small centroid motion the median aperture is optimal since it finds the “core” pixel array containing the signal. The median is also better than the mean at not being skewed by transient blips in the centroid position. The median aperture therefore robustly removes outlier cadences. When there is large motion however, the “core” aperture found with the median will clip too much of the flux far from the center of motion, while the 95% union aperture is better at accounting for the “smear” of the flux over the larger pixel array.

By ordering the pixels as described above, the number of possible apertures in the  $N$ -pixel mask is reduced from  $2^N - 1$  (all possible subsets of pixels minus the empty set) to  $N$ . With this ordering we can therefore afford to compute the much more costly estimate of Combined Differential Photometric Precision (CDPP, Jenkins et al., 2010) for each of the  $N$  candidate apertures to find the aperture with the lowest CDPP. This CDPP-optimized aperture is discussed in Subsection 7.4.2.

#### 7.4.1 Modeling Target Masks

The optimal aperture selection is determined by separating the flux contribution of the target star from that of other sources so that we can disentangle the numerator, or information, from the denominator, or noise, in the SNR. An accurate model of image formation enables this decomposition of background-subtracted pixel flux measurements into contributions from each point source in the instrument’s FOV. Our model consists of a scene description comprising the set of known sources  $\mathbf{s} = \{s_1, s_2, \dots, s_N\}$  with celestial coordinates  $(\alpha_n, \delta_n)$  and magnitudes, a model of background flux, a model of image motion on the focal plane and a model of the PRF.

We model the calibrated mean flux  $f$  at each pixel  $p$  during cadence  $c$  as a linear combination of PRF values for each source plus a background model,

$$f(p, c) = f_{bkgnd}(p, c) + \beta(c) + \sum_{n=1}^N F_n(c) \widehat{PRF}(p, x_n(c), y_n(c)). \quad (7.4)$$

The total flux  $F_n$  due to source  $n$  is distributed among pixels according to the normalized PRF, referred to here as  $\widehat{PRF}$ , which is obtained by evaluating the PRF over an 11-by-11 pixel grid centered at  $x, y$  and dividing pixel  $p$  by the sum of evaluated pixels. For more details on evaluating the PRF the reader is referred to (Bryson et al., 2010a). The real-valued CCD coordinates  $x_n(c)$  and  $y_n(c)$  are estimates of the mean photocenter of source  $s_n$  during cadence  $c$ . Centroid estimates are obtained from a polynomial model of image motion, which maps celestial coordinates to CCD coordinates,

$$x_n = P_x(\alpha_n, \delta_n) \quad (7.5)$$

$$y_n = P_y(\alpha_n, \delta_n). \quad (7.6)$$

Motion polynomials  $P_x$  and  $P_y$  for each channel and cadence are robustly fitted to images of a spatially distributed set of  $\approx 200$  bright, unsaturated and uncrowded sources (Twicken et al., 2010, see also Chapter 6). Background flux estimates  $f_{bkgrnd}$  are obtained from similarly-constructed 2-D polynomial models fitted to the grid of background pixels collected from each channel. The constant term  $\beta$  is included to account for any local bias in the background flux estimates for the set of pixels being modeled.

An independent model is fitted to each set of pixels comprising a target mask. The scene model for a given mask is constructed from entries in the KIC (Brown et al., 2011) UKIRT (Lawrence et al., 2007) and Ecliptic Plane Input Catalog (EPIC)<sup>4</sup> (Huber et al., 2016) and is initialized by ranking sources within 5 pixels from the edges of the target’s mask by their maximum expected SNR contribution to any pixel in the mask. Sources whose maximum SNR values fall below a threshold are discarded and up to 30 sources from the remaining set are admitted to the scene model in order of rank. Since certain regions of the FOV may be densely populated with dim stars, limiting the scene model in this way prevents wasting time and computing power on sources that do not make significant flux contributions.

We fit model parameters  $F_n(c)$ ,  $\beta_n(c)$ ,  $\alpha_n$ , and  $\delta_n$  to the observed data by minimizing the weighted sum of  $\chi^2$  residuals over all pixels and cadences in the model. Fine-tuning the source positions  $\{\alpha_n, \delta_n\}$  in the fit compensates for both catalog errors and local biases in the motion polynomials. Position fitting is enabled only for sources with centroids inside the target mask that contribute sufficient flux, defined by a threshold parameter. We only allow for position variations if the star is sufficiently bright *and* its expected centroid lies within the mask. If the predicted centroid falls outside the mask, then the data would only capture the “skirt” of the PRF and its position would be too poorly constrained for a fit to be reliable. The minimization is constrained such that source flux is non-negative and perturbations,  $\Delta_n$ , to catalog positions are smaller than 1.5 pixels. The purpose of the fitting of  $\Delta_n$  is to identify and correct errors in the catalog. But we have found in some cases the fitter will drift far off in R.A. and decl. and so the limit of 1.5 pixels (3.98 arcsec) is to inhibit these cases. A drift of greater than 4 arcsec is considered a poor fit since catalog errors are not expected to be any larger (Brown et al., 2011). Note that the 1.5 pixel limit is only to the catalog right ascension and declination  $\{\alpha_n, \delta_n\}$ . The motion polynomials can still allow for greater centroid motion than 1.5 pixels across the CCD, which is not uncommon for K2 data. Also note that, while fitting  $F_n$  and  $\beta$  is a linear problem, fitting source positions  $\{\alpha_n, \delta_n\}$  is inherently nonlinear. We address the two components of the fit separately in an iterative scheme. We use a Levenberg-Marquardt algorithm to optimize the perturbation of source positions, while the non-negative least squares method in (Kim et al., 2013) is used to fit the flux

<sup>4</sup>EPIC is hosted at MAST (<http://archive.stsci.edu/k2/epic/search.php>).

model of Equation 7.4 for a given perturbation,

$$\begin{aligned} & \text{minimize} && (w(\vec{\Delta}) + 1) \sum_{p=1}^P \sum_{c=1}^C \chi^2(p, c; \vec{F}(c), \beta(c), \vec{\Delta}) \\ & \text{subject to} && F_n(c) \geq 0, \Delta_n \leq \Delta_{max} \end{aligned} \quad (7.7)$$

and where

$$\chi^2(p, c) = \frac{(f_{\text{observed}}(p, c) - f_{\text{model}}(p, c))^2}{\sigma(p, c)}. \quad (7.8)$$

Measurement uncertainties,  $\sigma(p, c)$ , are obtained from CAL along with calibrated pixel flux,  $f_{\text{observed}}$ . The heuristic weighting function,  $w(\vec{\Delta})$ , in Equation 7.7 prevents source positions from converging on the same peak in the data by increasing as sources move toward one another from their catalog positions. Treating each source as a charged particle of like sign, we compute the potential energy of the source configuration defined by the catalog as well as that of the perturbed configuration (catalog +  $\vec{\Delta}$ ). If the energy of the perturbed configuration is greater than the catalog configuration then  $w(\vec{\Delta})$  is given the value of the difference (perturbed - catalog). Otherwise it is set to zero. By basing the weight on the energy *increase*, we penalize solutions in which multiple stars tend toward the same position without penalizing solutions that agree with the catalog. An illustration of the model fitting process is given in Figure 7.2.

The fit yields an initial light curve estimate for the target star as well as estimates of errors in its catalog position and magnitude. The light curve estimate constitutes the numerator of the SNR estimate in Equation 7.3 used for establishing the optimal aperture and pixel adding order. The quantities  $F_n(c)$ ,  $\alpha_n$ , and  $\delta_n$  can be used to identify errors in the *Kepler* KIC, UKIRT and EPIC Catalogs.

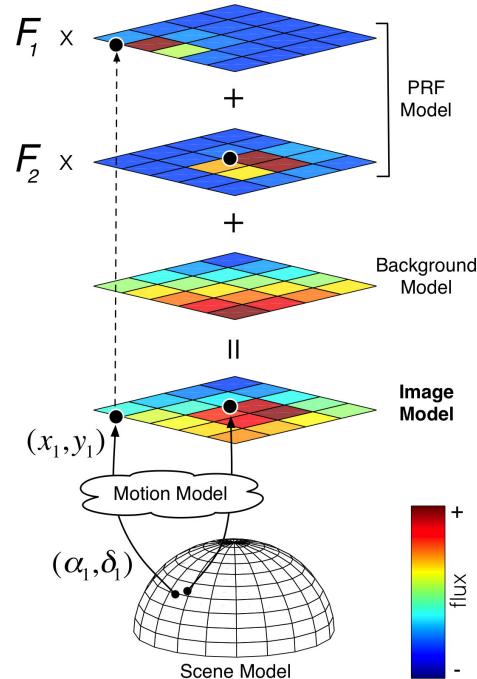
#### 7.4.2 Optimizing the Aperture for Photometric Precision

SNR is a very simple and fast calculation. One can easily calculate it for every pixel combination and find a pixel adding order in a tractable amount of time, even for large apertures. However, if planet transit finding is the primary science goal, as it is for the *Kepler* Mission, a better metric to optimize the aperture is Combined Differential Photometric Precision (CDPP). Simply speaking, CDPP is an estimate of how well a transit-like signal can be detected in a stellar light curve (Jenkins et al., 2010). A 6-hour CDPP of 100 part-per-million (ppm) means that a 100 ppm 6-hour transit signal has a detection statistic of  $1\sigma$ . CDPP is a significantly more sophisticated calculation than SNR, so we cannot afford to calculate it for a large number of candidate apertures. Fortunately, in Section 7.4 we have already described a pixel adding order per cadence. But a further step is first needed since CDPP is calculated on a flux time series, not on individual cadences. Therefore, a single “averaged” pixel adding order over the entire quarter is needed. This single “average” pixel adding order gives us a limited set of

apertures where we can compute CDPP. We calculate the “average” pixel adding order by taking the “mode”, or most frequently occurring values, of the pixel adding order per cadence, ensuring that every pixel is only included once in the averaged pixel adding order. Once we have this order, we begin with the center pixel (first pixel in the pixel adding order) and generate a simple aperture light curve. CDPP is then calculated for each sequentially larger aperture using the averaged pixel adding order.

The calculation of CDPP is explained in detail in Jenkins et al. (2010, see also Chapter 9). Calculating CDPP for all 14 pulse durations used in the planet transit search would be prohibitively expensive here. The performance benchmark for *Kepler* is a 6-hr CDPP on a 12th magnitude dwarf star. We therefore restrict ourselves to computing a 6-hour transit duration CDPP. Also considered is that the total flux in the aperture changes as we increase the aperture size, which will directly change the absolute CDPP values. We therefore median normalize the light curves to place each aperture on equal footing. The CDPP thus computed is a time series giving the detection statistic for each cadence. We take a robust root-mean-square (rms) value of this time series to obtain a single CDPP value as a function of aperture size.

CDPP is dangerous to use by itself to find the optimal aperture. If, for example, a background object is brighter than the target, then the CDPP of an aperture centered on the background object will very likely be less than one centered on the target, and the aperture will grow to engulf the brighter background object. The SNR, on the other hand, is very robust against background contamination, the numerator in the SNR being purely from the modeled target star. Any contamination will only decrease the SNR value. A combination of the two metrics can therefore simultaneously exclude background objects and



**Figure 7.2** A single-cadence model of an actual *Kepler* target pixel mask. The image model is constructed from a background model, a normalized PRF model evaluated for each source, and estimates  $F_n$  of the total electron flux contributed by each source. The celestial coordinates  $(\alpha_n, \delta_n)$  of each source are mapped to CCD coordinates  $(x_n, y_n)$  by a polynomial model of image motion. Note that in this example the centroid of source  $s_1$  lies slightly outside the mask, but it still contributes significant flux to pixels inside the mask. Also note that the color spaces of the component images have been stretched to aid visualization. From Figure 1 in Smith et al. (2016).

optimize for planet detection. We combine the two and find  $k$  to minimize

$$\widetilde{\text{CDPP}}_k - (\text{SNR}_k - \overline{\text{SNR}}) \frac{\sigma_{\text{SNR}}}{\sigma_{\text{CDPP}}}, \quad (7.9)$$

where  $\widetilde{\text{CDPP}}_k$  is a smoothed CDPP curve for an aperture comprising  $k$  pixels. The smoothing is necessary since the CDPP measurement has more scatter than the SNR curve, which is typically very smooth.  $\sigma_{\text{SNR}}$  and  $\sigma_{\text{CDPP}}$  are the variance in the SNR and CDPP curves, respectively, and the ratio of these scales the SNR curve to the same range as the CDPP curve so that they are given equal weight in the optimization.  $\overline{\text{SNR}}$  is the median SNR value over aperture sizes  $k$ . The minimum of the above curve is chosen and this is the CDPP-optimized aperture.

## 7.5 Selecting the Best Aperture using CDPP and Logistic Regression

We have now found four apertures:

1. Pure PRF image model-derived (Method #1)
2. Real pixel data-derived SNR-optimized Median
3. Real pixel data-derived SNR-optimized Union
4. Real pixel data-derived CDPP-optimized.

Each of these apertures can yield better performance for a subset of *Kepler* targets. Since transit signal detection is the primary goal of the *Kepler* Mission, we principally select the optimal aperture based on CDPP. An estimate of the uncertainty in the CDPP measurement is used to bias the selection to just apertures 2 – 4. This is because we believe an aperture selected using flight data will be superior in most cases to a purely model-based aperture. So if we are within the uncertainty in the measurement, we will pick the Method #2 aperture. However, in certain cases, the data-derived aperture can fall into a local minimum during the optimization procedure and not identify the true optimal aperture, hence the model-based aperture (Method #1) is still considered, which has more consistent and robust behavior (by virtue of it being based on a synthetic model).

For about 0.5% of targets, Method #2 has been shown to be problematic, in spite of it having a lower CDPP. CDPP is agnostic to whether or not the proper target is chosen. For example, if one of the four apertures incorrectly centers on a background object that is brighter than the target then the CDPP metric could select that aperture. Method #1, on the other hand, is more robust to these cases, but at the expense of overall poorer apertures. So for the 0.5% of targets we identify, reverting to Method #1 is a move to revert to the more conservative aperture. A logistic regressive heuristic model was developed to help identify these corner cases. We identified several key metrics that can be used in aggregate to identify the bad corner cases including:

1. Net change in flux in the aperture between Method #1 and Method #2,
2. Fractional change in aperture between Method #1 and Method #2,
3. Fractional change in flux between Method #1 and Method #2,
4. Ratio of total mask used in the Method #2 aperture, etc.
5. Fraction of total mask used in the Method #2 aperture,
6. etc.

Any one of these predictors is not necessarily sufficient to identify poor performance. Logistic regression (see James et al., 2014), was therefore used to identify the appropriate combination of predictors. To find the proper regressive model, a training set is needed to “train” the model. Each of the metrics was calculated for all targets on *Kepler* CCD module channel output 2.1 for quarter 13, 1358 targets in total. The targets were ranked by response to each of the metrics. Then for each target in order of rank, the diagnostic information generated by the aperture finding algorithms (an example of which is shown in Figure 7.4 and Figure 7.5) was examined to identify poor aperture selection in Method #2. Targets continued to be tallied down to lower predictor response until nearly all remaining Method #2 apertures appeared to be good. This resulted in a training set of 100 targets, with 10 having incorrectly chosen one of the Method #2 apertures. Multinomial Logistic Regression was then used to model a response  $p$  on the predictors:

$$p(\mathbf{X}) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_N X_N}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_N X_N}} \quad (7.10)$$

where  $\mathbf{X} = (X_1, \dots, X_N)$  are the  $N$  predictors and  $\beta = (\beta_0, \dots, \beta_N)$  are the coefficients plus intercept  $\beta_0$ . The coefficients were fitted using the maximum likelihood, and the p-values for the identified coefficients were then used to rank the suitability of each of the predictors. After a couple iterations of dimensional reduction, we found that the three predictors, 1, 3 and 4 from above, are most effective at identifying poor aperture selection. The other predictors, such as the fractional change in aperture, were found to be poor predictors and we zeroed their coefficients,  $\beta_2$ , etc. This zeroing resulted in better separation in the prediction value  $p(\mathbf{X})$  between good and bad apertures<sup>5</sup>. Several other predictors were also tried but not listed above, as none were found to enhance the separability of good and bad apertures.

The prediction  $p(\mathbf{X})$  is a real function spanning  $[0, 1]$  so we must determine the prediction discrimination threshold for poor apertures,  $\tau$ . The threshold value that minimizes false negatives was chosen, the logic being whenever an aperture found by Method #2 is deemed poor, we error conservatively by falling back on

<sup>5</sup>This observation emphasizes the common adage that more variables in a model does not necessarily produce a better model!

**Table 7.1** Table of Confusion for identifying poor aperture selection.

True Positive	False Negative
22	3
False Positive	True Negative
18	3891

the more robust Method #1 aperture. Once our predictive model was trained, we used two different CCD channels, 7.3 and 13.2, to test the model (3934 targets in total). These two channels have different image characteristics than the training channel, 2.1, and so they are good channels with which to test the robustness of the method. The three predictors and coefficients were evaluated in Equation 7.10 to return a prediction value, and when  $p(\mathbf{X}) > \tau$  the target aperture was flagged as poorly chosen and we reverted to Method #1. The table of confusion for the test data is given in Table 7.1. We see that the total number of truly poor apertures is  $\frac{25}{3934} \approx 0.5\%$  and we correctly identify 22 of those, leaving only 3 or about 0.05% not identified. The 18 false positives are not considered problematic because, as explained above, we choose to error conservatively. This logistic regressive method can therefore correctly identify 90% of the small number of poor apertures, reducing the total number of poor apertures down to only 0.05%, which translates to only about 1 target per CCD channel, or  $\sim 75$  total out of the  $\sim 165,000$  targets processed by *Kepler*.

## 7.6 Per Cadence Flux Fraction and Crowding Metric

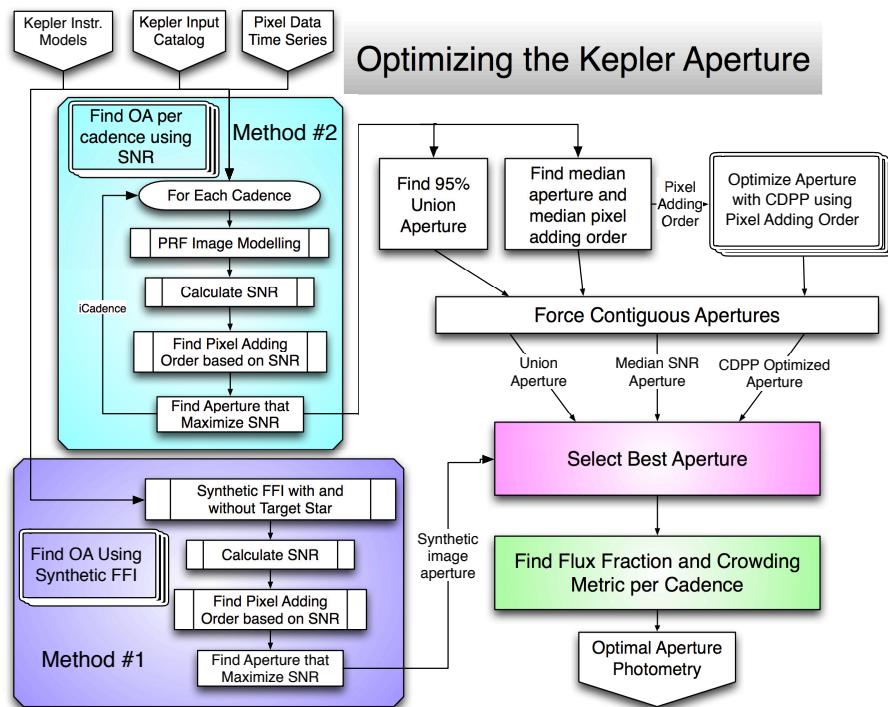
No aperture will collect all the flux from any one object, the non-finite PSF insists that some flux will spill over any finite aperture. We therefore calculate a Flux Fraction in Aperture, or *Flux Fraction* for short, to quantify the fraction of target flux in the photometric aperture. Likewise, background objects will inevitably spill into the aperture to some degree. The *Crowding Metric* gives the fraction of flux in the aperture that is due to the target star. Prior to the introduction of PA-COA in SOC 9.3, the mission was limited to a single Flux Fraction and Crowding Metric value per quarter. Since a target image model can be computed for every cadence, a Flux Fraction and Crowding Metric can now be computed per cadence. These per cadence values can be used to further reduce systematic trends in the data. However, we have not yet incorporated this information into the systematic error removal in PDC (see Chapter 8).

## 7.7 Saturated Pixels

Very bright targets result in CCD pixel saturation. For saturated targets the image modeling in Subsection 7.4.1 does not function properly because the PRF model does not account for saturation. We therefore always choose Method #1 for these

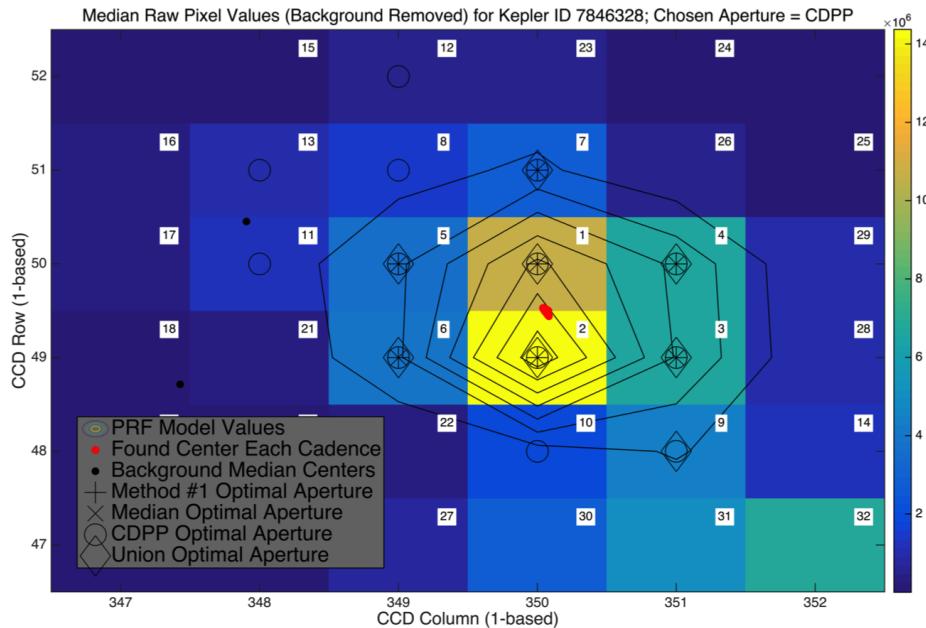
targets and use the purely synthetic image model. Saturated charge is spilled along columns, with the fraction of charge spilled up and down determined by the saturation model mentioned in Section 7.3. The pixel values are set to the saturation value in the synthetic image created in Method #1, and the aperture is computed. Ground tests indicate that the well depth and saturation spill direction varies across the focal plane (Van Cleve & Caldwell, 2016), including variation within a channel. Therefore a buffer is added to the optimal aperture for saturated targets. This buffer is applied equally up and down the column and is added to the target's optimal aperture, but the pixel values themselves are not changed. If the buffer size is smaller than the optimal aperture in either direction, then the buffer size is ignored in that direction. In principle, there is no reason the saturation model cannot be used in the image modeling in Subsection 7.4.1 and future improvements could incorporate this model.

## 7.8 Summary and Example of Performance



**Figure 7.3** Flow chart showing the steps in finding optimal apertures in *Kepler* data. From Figure 2 in Smith et al. (2016).

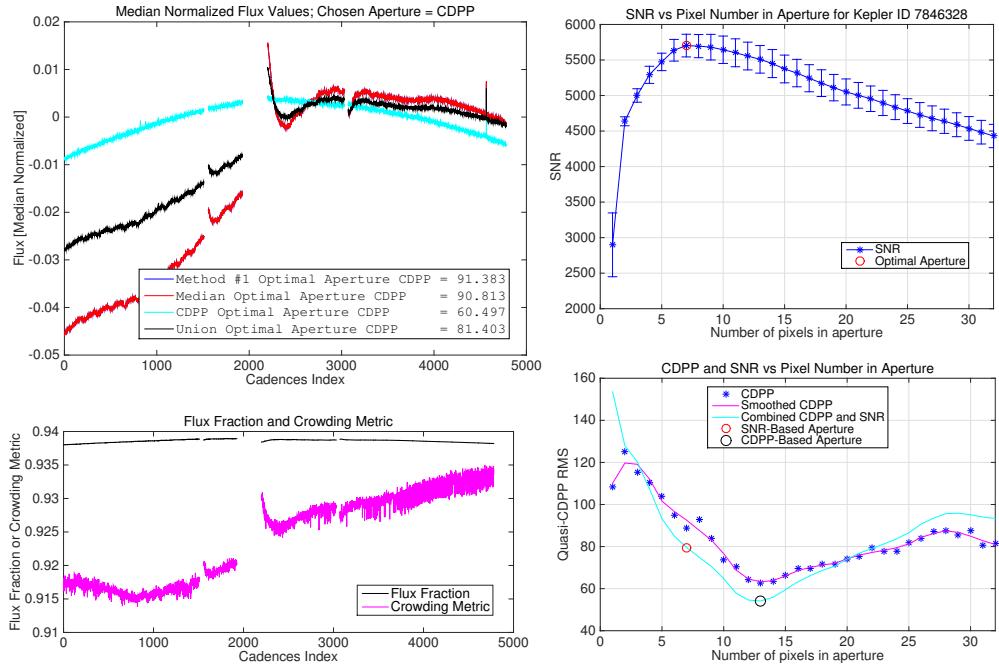
In Figure 7.3 we see a general flow chart of all the steps in aperture selection discussed in this paper. The purple box shows Method #1 and the aquamarine box shows Method #2, which loops over all cadences. We find four different apertures in total, force contiguity of the apertures and finally select the best. Per cadence Flux Fractions and Crowding Metrics are then computed for the chosen



**Figure 7.4** The mask scene, fitted PRF image and four found apertures for *Kepler* ID 7846328. The color bar is in units of electrons per second. From Figure 3 in Smith et al. (2016).

aperture. In practice, for *Kepler* data, we do not actually loop over all cadences but just every tenth cadence. We found no degradation in performance and yet achieved a nearly factor of ten decrease in processing time, allowing the pipeline to continue to finish in a timely manner. The image motion in *Kepler* data is a small fraction of a pixel per cadence and so every tenth cadence is sufficient to capture the full motion. As an example, Figure 7.4 shows the four found apertures for *Kepler* ID 7846328. The pixel image colors give the background-removed pixel data in this target’s mask. The found target center on each cadence is plotted as red dots (the centroid motion is not large for this example) and the median center for each background object is a black dot. In this example we see two dim background objects with black dots and one bright background object with a center just off the mask. The four found apertures are labelled as crosses, x’s, circles and diamonds as referenced in the legend. The contour lines show isophotes of the target object’s image model, which is used as the numerator in Equation 7.3. The pixel data (with background NOT removed) are used in the denominator of the SNR. The aperture selection process is illustrated for this same target in Figure 7.5. The upper right plot shows the SNR versus aperture size using the average pixel adding order with a maximum clearly present. Note that the actual SNR-optimized aperture is found for every cadence, but here we just plot the average SNR over all cadences. The error bars represent the standard deviation in the SNR curves over all cadences. This target has little motion or stellar variability so the SNR curve is very similar for all cadences. The lower right plot shows calculated CDPP versus the number of pixels in the aperture in the average pixel adding order as blue dots and the smoothed CDPP curve as a magenta

line. The cyan curve is the combined CDPP and SNR curve as in Equation 7.9. Combining the CDPP with SNR allows for a single minimum to be more pro-

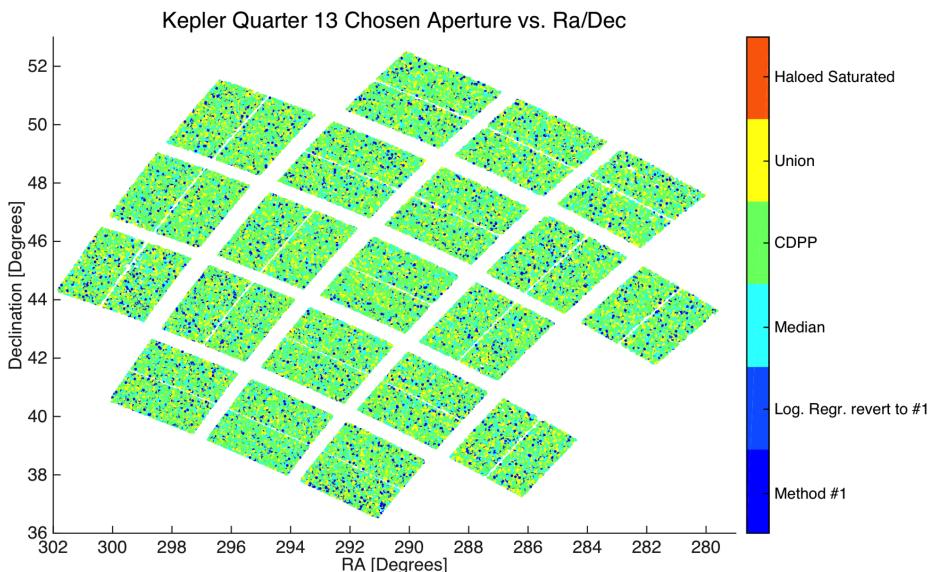


**Figure 7.5** The aperture fitting diagnostic figures for *Kepler* ID 7846328. From Figure 4 in Smith et al. (2016).

nounced. The magenta curve, on the other hand, has two potential local minima near zero and 32 pixels. Neither would be a proper aperture for this target. The dip in the magenta CDPP curve beginning at pixel index 30 is due to the aperture engulfing the bright background object as shown by the pixels labeled “30–32” in Figure 7.4. If the pixel mask for this target was even larger and the CDPP curve extended out to cover the entire bright background object then the minimum in the CDPP curve would include both objects, as occurs in many other target examples. The upper left plot shows the photometric light curve for each of the four apertures. In this case the Method #1 and median SNR-optimized apertures contain the same pixels (as also seen in Figure 7.4) and so the blue curve lies directly under the red. The legend contains the calculated rms CDPP for each of these four light curves and the CDPP-optimized light curve (cyan curve) is the clear winner at 60 ppm, versus 91 ppm for the Method #1 aperture. We can also see that the cyan curve has dramatically reduced instrumental systematics versus the other curves and so it is the clear superior aperture and light curve. Systematic features in the data comparable in length and size to transits will obscure transit signals and therefore increase CDPP. It is therefore reasonable that the CDPP optimized light curve will also have minimized systematics. The three-day reaction wheel heater cycle visible in the black and red curves is a good example of a systematic signal that can obscure a transit signal. One may suspect that the

CDPP-optimized aperture will always win the CDPP test. If we were optimizing the aperture purely on CDPP then this would be the case. But as discussed above, finding a global CDPP minimum would require  $2^N - 1$  evaluations of CDPP for a pixel mask of size  $N$ . This is an insurmountable computational endeavour for an automated pipeline. We therefore must reduce the number of CDPP computations, which we do by finding the average pixel adding order. Doing so risks not finding the true optimized adding order and therefore the CDPP-optimized aperture can drift off in aperture space and not find the best aperture. We also have to take care not to engulf background objects, further complicating the CDPP optimization. The heuristic logistic regressive model will also revert to the Method #1 aperture in some cases. Considering for these caveats the actual aperture used on all targets is a combination of the four found apertures. For quarter 13, over 166,791 targets, the breakdown for aperture selection is

- Pure PRF image model derived = 7.6%
- Real pixel data derived SNR-optimized Median = 21.3%
- Real pixel data derived SNR-optimized Union = 22.5%
- Real pixel data derived CDPP-optimized = 48.6%.



**Figure 7.6** The chosen aperture for *Kepler* Quarter 13 over the Field of View. The aperture selection is evenly distributed over the FOV with a strong preference for the CDPP-optimized aperture. From Figure 5 in Smith et al. (2016).

Of the 7.6% that use Method #1, 0.9% are due to the logistic regressive heuristic model identifying when the Method #2 aperture is poor despite it having the lowest CDPP. Note also that 21.3% chose the SNR-optimized median aperture over the CDPP-optimized aperture. This is again due to the CDPP-optimized

aperture not being optimized solely on CDPP, and also because many times the SNR- and CDPP-optimized apertures have very similar CDPP values. The fractional change in aperture between Method #1 and Method #2 does have an annular shape across the FOV, where targets further away from the ring of best focus require more correction. This is in agreement with expectations. We rely on a static PRF model that was derived from data acquired in a particular state of focus during commissioning in May 2009. Our image models are therefore generally better in regions that more closely match that state. A graphical representation of the selected aperture over the FOV is shown in Figure 7.6. We see the dominance of the CDPP-optimized aperture selection and an even distribution over the FOV for all aperture types but with a slight preference for Method #2 apertures near the center of the FOV. There are no “Haloed Saturated” apertures, this aperture type is reserved for K2 data as discussed in Section 7.9.

The full *Kepler* data set of all 18 Quarters of data has been reprocessed using the newer aperture selection method discussed in this paper. Figure 7.7 shows a histogram of the overall improvement in CDPP for simple aperture photometry light curves using the newer method (the old method being using solely Method #1). The revised apertures reduce CDPP for the vast majority of targets and quarters, with 77% showing an improvement, 13% showing some degradation, and 10% showing no change. The majority of those showing no change is due to the algorithm choosing Method #1. In 14% of cases CDPP is reduced by 10% or more. The next component in the *Kepler* Science Data Processing Pipeline is PDC (Stumpe et al., 2012, see also Chapter 8) so light curves generated with simple aperture photometry in PA are not the final light curves generated by the *Kepler* Mission. In PDC we remove systematic trends in the data and so the CDPP values generated in PA are much larger than those seen by the transiting planet search algorithm. But with the improvements in this paper, PDC now begins with lower noise light curves than before. The resultant PDC light curves now have even lower CDPP, which directly translates into better sensitivity to transit detection in TPS. We have shown that the decrease in CDPP does indeed propagate through PDC.

## 7.9 Application to K2 Data

In the K2 Mission (Howell et al., 2014), periodic thruster firings are used to compensate for the loss of two of the four reaction wheels. These firings occur up to once every 6 hours, or every 12th long cadence, and the resulting barrel axis roll motion can be well over 4" for the outer CCD channels. The roll drift and thruster firing repeats produce a characteristic “sawtooth” pattern in uncorrected light curves. Method #1 as implemented is unable to account for the K2 roll motion and so the Method #1 apertures are typically of low quality for the higher motion targets. Method #2, in contrast, can be directly applied to the K2 Mission data and there are very few changes to the algorithm. To properly account for

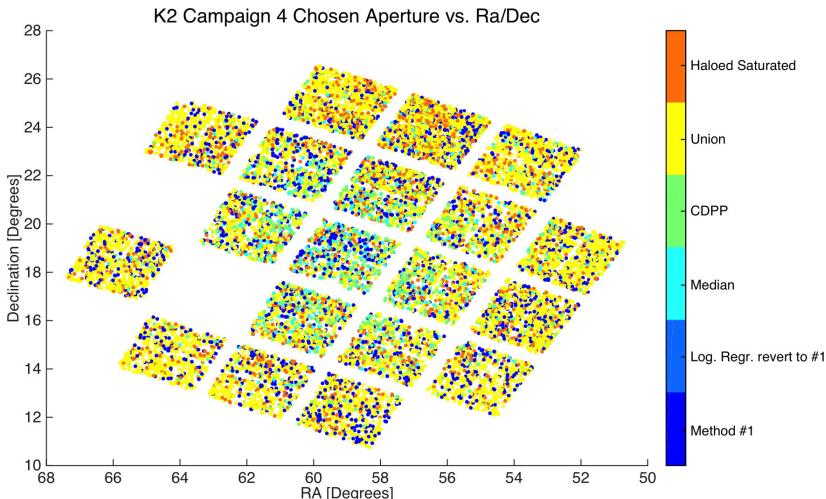
the large, fast motion we must find the aperture on every cadence. There are no changes in how the optimal aperture is selected but due to this large motion the union aperture is most often found to have the lowest CDPP as shown in the following aperture selection breakdown for Campaign 4 with over 17,278 targets:

- Pure PRF image model derived = 34.4%
- Real pixel data derived SNR-optimized Median = 4.3%
- Real pixel data derived SNR-optimized Union = 53.3%
- Real pixel data derived CDPP-optimized = 8.0%.

Of the 34.4% that chose Method #1, 19.4% are saturated or custom targets where Method #1 is always chosen. This leaves only 15.0% where the Method #1 aperture was considered better than the ones selected by Method #2. A graphical representation of the selected apertures over the FOV is shown in Figure 7.8. We see the annular dependence where targets further from the boresight, and where the roll motion is larger, strongly prefer the union aperture. Near the center of the FOV, where the roll motion is smaller, the Median and CDPP apertures are more preferentially chosen. For saturated targets and those with large roll motion, Method #1 will not always collect all relevant pixels. We therefore add a two pixel halo around each Method #1-derived saturated target aperture, these targets are labeled as “Haloed Saturated” in the figure and are evenly distributed. Notice that the Method #1 apertures are clustered. The vast majority of these are custom targets where the Method #1 aperture is always chosen and custom targets tend to be clustered in regions of interest to specific Guest Observer Office funded projects<sup>6</sup>.

To illustrate the performance on K2 data, Figure 7.9 and Figure 7.10 give the pixel scene, apertures and diagnostic curves for EPIC ID 204115036 processed during Campaign 2 on module output 2.3. These two figures are directly analogous to the example figures given above for *Kepler* data. Here, we clearly see the much larger motion (red centroid dots) in the K2 Mission. The Method #1 aperture in Figure 7.9 does not properly account for the large roll motion and therefore finds an aperture (the plus symbols) that is slightly off the target centroid. The Method #2 median and CDPP-optimized apertures (the x's and circles, respectively) do find the proper center but do not fully account for the motion. The union aperture, however, is both properly centered and accounts for the full motion of the target centroid. Three bright background objects are very near to the target. Care must be taken that these objects do not contaminate the target's optimal aperture. In Figure 7.10 we see the diagnostic curves. Note that the “average” SNR curve no longer peaks at the SNR-optimized median aperture size (i.e., the red circle in the upper right plot is not on the blue curve peak). This

<sup>6</sup>See <http://keplerscience.arc.nasa.gov> for details on the Guest Observer program.

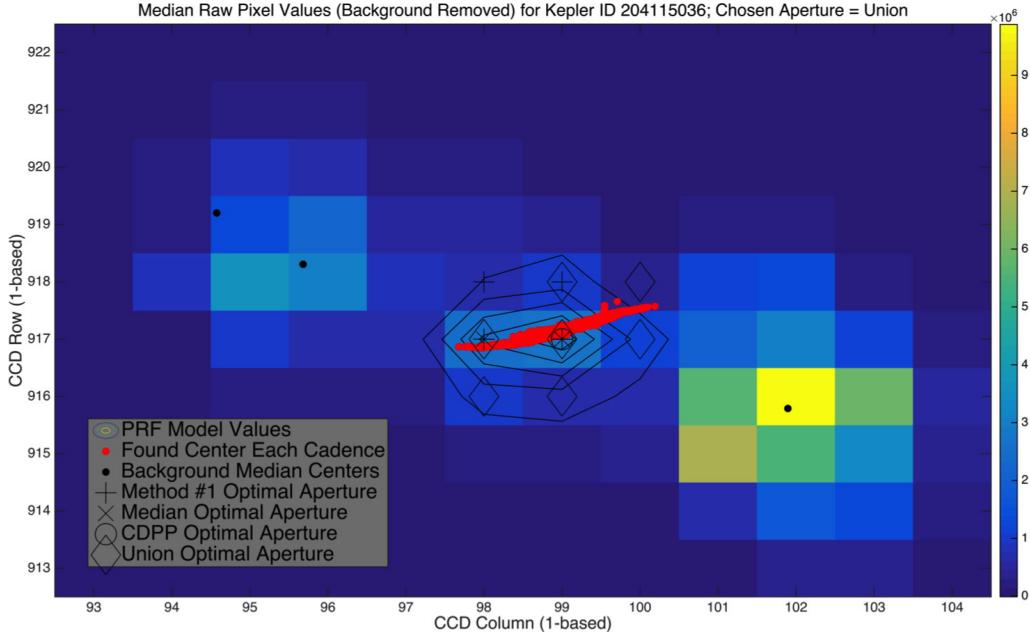


**Figure 7.8** The chosen aperture for K2 Campaign 4 over the FOV. The annular dependence is due to the large roll motion about the instrument barrel axis. From Figure 7 in Smith et al. (2016).

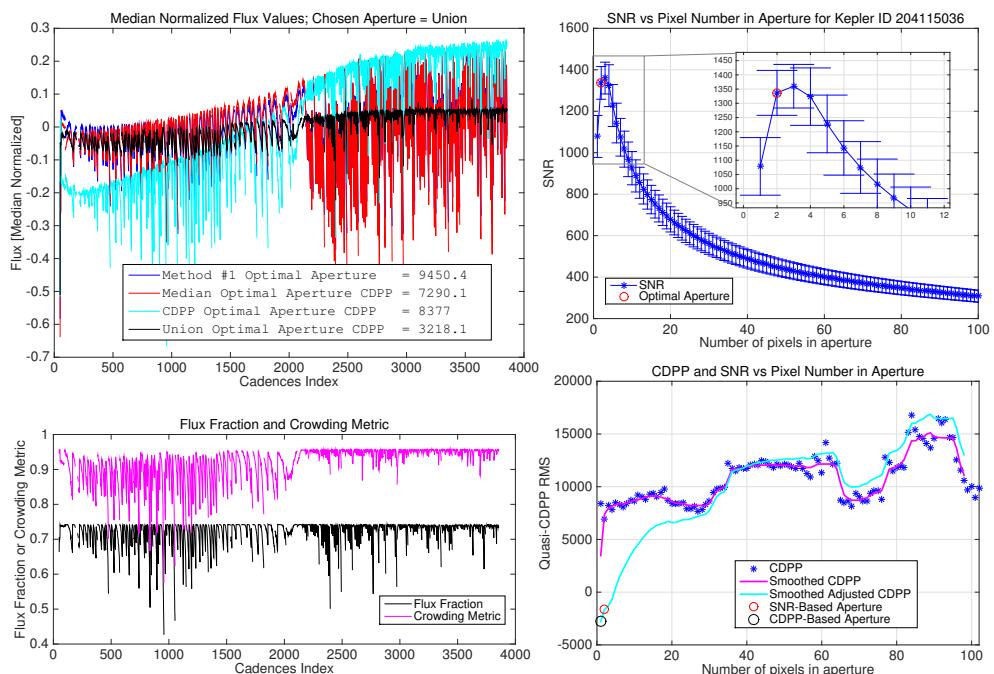
is indicative of high centroid motion and that the SNR-optimized median aperture is changing over time. Therefore the “averaged” SNR-optimized curve will not account very well for the motion nor find a good optimal aperture. This can also be seen in the large error bars that represent the standard deviation in the SNR curves over all cadences. The spread is large compared to the sharp peak in the curve. We also see in the lower right plot that the (magenta) CDPP curve has minima as we engulf the background objects, whereas the combined CDPP and SNR curve does not (cyan curve). But neither the median nor the CDPP-optimized aperture is the best aperture. The upper left plot shows that the union aperture has, by far, the lowest CDPP and is the best choice. The large image motion relative to the optimal aperture results in the uncorrected light curves, flux fraction and crowding metric time series showing a strong “sawtooth” behavior. We could increase the aperture size until no sawtooth is present, but doing so would cause the SNR to decrease and the broad-spectrum noise to increase to unacceptable levels; we would simply be trading a distinct noise source for broad spectrum noise. The distinct sawtooth signal can be removed in PDC, whereas broad spectrum noise is very difficult to remove. We therefore are best served by minimizing the noise at the expense of a stronger sawtooth signal.

The above discussion invites the natural conclusion that adaptive apertures that adjust on every cadence might be optimal for K2 data. The authors do not disagree with this observation. Unfortunately, the K2 Science Data Processing Pipeline is an adaptation of the *Kepler* Science Data Processing Pipeline, which was committed to a single fixed optimal aperture per quarter. The time and resources were not available to modify the pipeline to allow for adaptive apertures. Given this constraint we have found smaller apertures to be generally better.

In Subsection 7.4.1 we mentioned that the fitted image model parameters can be used to identify errors in the stellar catalogs. Two examples of this were found

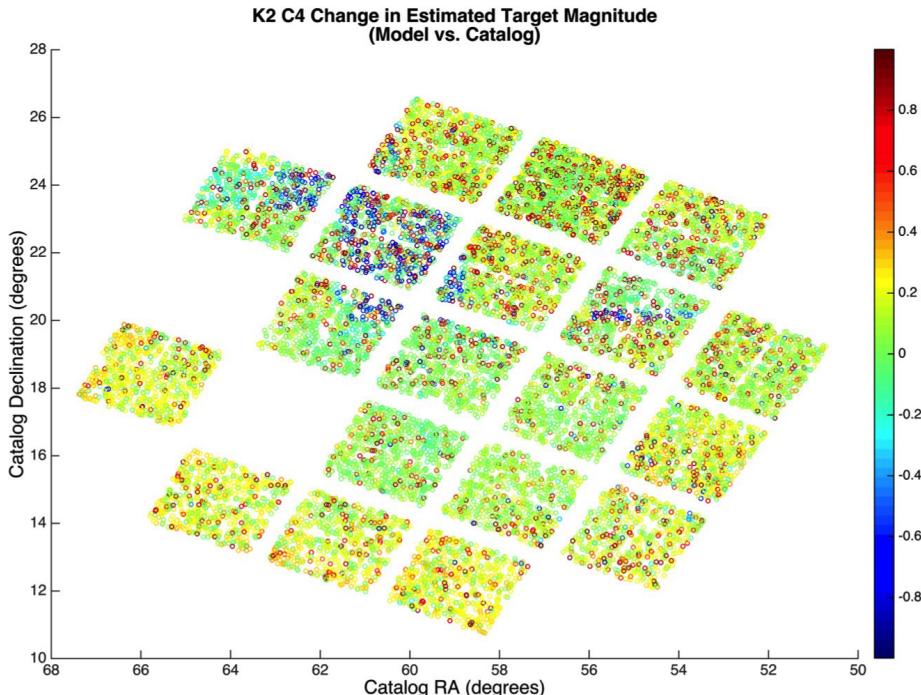


**Figure 7.9** The mask scene, fitted PRF image and four found apertures for EPIC ID 204115036. From Figure 8 in Smith et al. (2016).



**Figure 7.10** The aperture fitting diagnostic figures for EPIC ID 204115036. From Figure 9 in Smith et al. (2016).

in K2 Campaign 4 processing. There is a group of targets whose measured flux is more than twice that expected from their EPIC magnitudes. Figure 7.11 shows that these targets fall into spatial groups that are aligned with right ascension and declination (shown as blue markers), rather than focal plane coordinates, strongly indicating that the cause of this anomaly is catalog errors. The source of this error is unknown and is not correlated with any particular target type. The other error identified in the figure is the scatter of red markers indicating targets whose brightness is overestimated. These targets are strongly correlated with “JHK” and “J” stars which were discovered to be K/M dwarfs that are not well represented in the data used to create the conversion from JHK/J to Kepler magnitude (Howell et al., 2012). Method #2 attempts to correct for these errors and find a true optimal aperture, but the PRF modeling can only go so far and so identifying these catalog errors is of benefit to the mission. Star catalogs generated for future campaigns will try to account for and fix these errors.



**Figure 7.11** All C4 target stars plotted in celestial coordinates, colored by their magnitude inferred from their observed flux minus their *Kepler* magnitude from the EPIC. There are two square-like regions and a line of blue markers, indicating stars whose inferred *Kepler* magnitude is about a magnitude smaller than their catalog magnitude, indicating that these stars are about a magnitude brighter than expected. The red markers are consistent with the population of “JHK” or “J” stars whose brightness is overestimated. From Figure 10 in Smith et al. (2016).

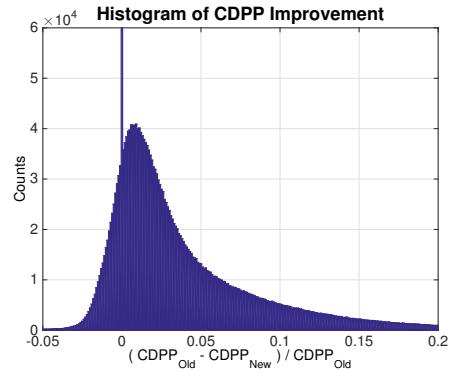
## 7.10 Conclusions

The above described optimal aperture finding method has been shown to provide superior photometry for *Kepler* data as summarized in Figure 7.7. We can now

also find a per cadence Flux Fraction in Aperture and Crowding Metric, providing a potentially superior systematic error removal; however, the latter was not implemented in the pipeline. The method further allows us to identify errors in the *Kepler* and K2 input catalogs.

The new approach has proven robust at finding apertures in K2 data and helping mitigate the larger motion-induced systematics in the photometry. The older Method #1 does not perform well with the larger motion and so the new method is absolutely critical for extracting high-quality photometry with K2. Dynamic and moving apertures could potentially provide even better photometry but the mission was not provided the resources to implement this further improvement. The next processing component in the *Kepler* Pipeline is PDC, which has been modified for use with K2 data (Van Cleve et al., 2016). It has been shown to remove up to 99% of the “sawtooth” systematic pattern in the data. Nevertheless, improvements to the PDC component of the K2 Pipeline is probably the area where the most improvement in photometry could be achieved.

As with any working method, improvements could still be made. One relates to how we find a single aperture from the per cadence optimal apertures in Equation 7.3 of Section 7.4. As of now we find two apertures: 1) a 50% union, or median aperture, and 2) a 95% union. Instead, we could allow the union percentile to be a parameter that we optimize versus SNR or CDPP. This would not require us to first find an average pixel adding order, which introduces an added approximation and sometimes, an awkward pixel order. Another potential improvement is to identify and fit for uncatalogued background objects. As of now, we use the KIC, UKIRT and EPIC  $\alpha$ ,  $\delta$ , and magnitude with the motion polynomials (Equation 7.6) of each object to model the scene and image motion. We do know dim, uncatalogued background objects do exist and can contaminate the scene fitting. A method to auto-find these objects would aid in more complete scene modeling. A final improvement could be a better characterization of the intra-pixel response variations. The PRF model obtained during commissioning has been proven to perform well for *Kepler*; however, we do know it does not capture all variations in the flux in the presence of motion and thermal variations. Commissioning and PRF modeling is critical to any CCD based ob-



**Figure 7.7** Histogram of fractional CDPP improvement when comparing light curves produced from the new aperture method with those produced from the original Method #1. Positive values indicate improved photometric precision. This histogram is over all *Kepler* targets and each target is counted separately for each quarter. That is, the sum of the histogram is (number of targets)  $\times$  (number of quarters). From Figure 6 in Smith et al. (2016).

servations, and the limits of the PRF model are probably a limiting factor in the performance of our method.

## Bibliography

- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A., 2011. “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” AJ, 142, 112
- Bryson, S. T. 2008. “Target Optimal Aperture Selection,” Tech. Rep. KADN-26108, NASA KPO@Ames Design Note
- Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al., 2010. “The Kepler Pixel Response Function,” ApJL, 713, L97
- Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010b. “Selecting Pixels for Kepler Downlink,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401D
- Howell, S. B., Rowe, J. F., Bryson, S. T., et al., 2012. “Kepler-21b: A 1.6  $R_{Earth}$  Planet Transiting the Bright Oscillating F Subgiant Star HD 179070,” ApJ, 746, 123
- Howell, S. B., Sobeck, C., Haas, M., et al., 2014. “The K2 Mission: Characterization and Early Results,” PASP, 126, 398
- Huber, D., Bryson, S. T., Haas, M. R., et al., 2016. “The K2 Ecliptic Plane Input Catalog (EPIC) and Stellar Classifications of 138,600 Targets in Campaigns 1-8,” ApJS, 224, 2
- James, G., Witten, D., Hastie, T., & Tibshirani, R. 2014. An Introduction to Statistical Learning (Springer)
- Jenkins, J. M., Caldwell, D. A., & Gilliland, R. L. 2004. KeplerAlgorithm Theoretical Basis Document: KSOC-21008 (Moffett Field, CA: NASA Ames Research Center)
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010. “Transiting Planet Search in the Kepler Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D
- Kim, D., Sra, S., & Dhillon, I. S., 2013. “A Non-Monotonic Method for Large-scale Non-negative Least Squares,” Optimization Methods Software, 28, 1012
- Lawrence, A., Warren, S. J., Almaini, O., et al., 2007. “The UKIRT Infrared Deep Sky Survey (UKIDSS),” MNRAS, 379, 1599
- Smith, J. C., Morris, R. L., Jenkins, J. M., et al., 2016. “Finding Optimal Aper-tures in Kepler Data,” PASP, 128, 124501

- Stumpe, M. C., Smith, J. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning I – Architecture and Algorithms for Error Correction in Kepler Light Curves,” PASP, 124, 985
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Twicken, J. D., Clarke, B. D., Bryson, S. T., et al. 2010. “Photometric Analysis in the Kepler Science Operations Center Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774023
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)
- Van Cleve, J. E., Howell, S. B., Smith, J. C., et al., 2016. “That’s How We Roll: The NASA K2 Mission Science Products and Their Performance Metrics,” PASP, 128, 075002

# CHAPTER 8

---

## PRESEARCH DATA CONDITIONING

JEFFREY C. SMITH<sup>1</sup>, MARTIN C. STUMPE<sup>1</sup>, JON M. JENKINS<sup>2</sup>, JEFFREY E. VAN CLEVE<sup>1</sup>, FORREST R. GIROUARD<sup>3</sup>, JEFFERY J. KOLODZIEJCZAK<sup>4</sup>, SEAN D. McCauliff<sup>5</sup>, ROBERT L. MORRIS<sup>1</sup>, AND JOSEPH D. TWICKEN<sup>1</sup>

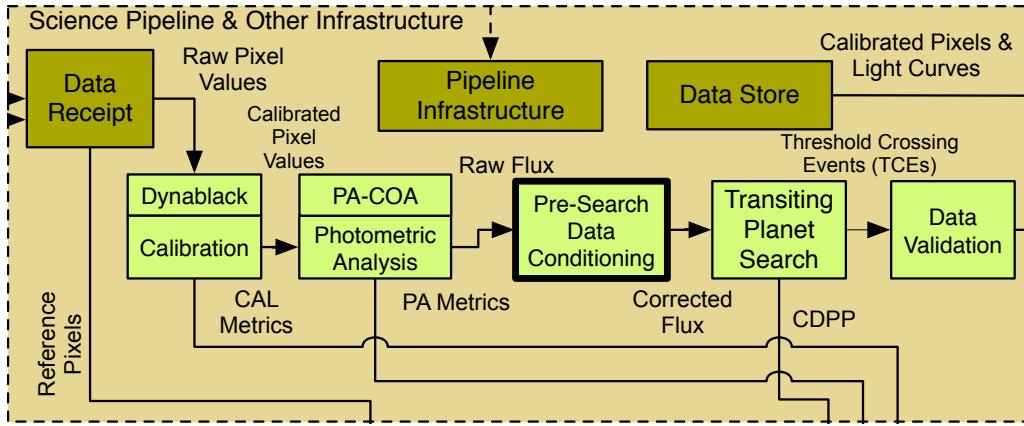
<sup>1</sup>The SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035, <sup>2</sup>NASA Ames Research Center, Moffett Field, CA 94035, <sup>3</sup>Logyx, LLC/NASA Ames Research Center, Moffett Field, CA 94035, <sup>4</sup>NASA Marshall Space Flight Center, Huntsville, AL 35808 <sup>5</sup>Wyle Labs/NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** Kepler provides light curves for  $\sim$ 200,000 stars with unprecedented precision. However, the raw data as they come from the spacecraft contain significant systematic and stochastic errors. These errors, which include discontinuities, systematic trends, and outliers, obscure the astrophysical signals in the light curves. Correcting these errors is the task of the Presearch Data Conditioning (PDC) module of the Kepler data science pipeline. The completely new noise and stellar variability regime observed in Kepler data poses a significant problem to standard cotrending methods. Variable stars are often of particular astrophysical interest, and the preservation of their signals is of significant importance to the astrophysical community. PDC first utilizes an overcomplete discrete wavelet transform, dividing each light curve into multiple channels, or bands, thereby allowing for a good separation of characteristic signals and systematics. The light curves in each band are then corrected utilizing a Bayesian maximum *a posteriori* (MAP) approach, where a subset of highly correlated and quiet stars is used to generate a cotrending basis vector set, which is in turn used to establish a range of “reasonable” robust fit parameters. These robust fit parameters are then used to generate a Bayesian prior and a Bayesian posterior probability distribution function (PDF) which, when maximized, finds the best fit that simultaneously removes systematic effects while reducing the signal distortion and noise injection that commonly afflicts simple least-squares (LS) fitting. A numerical and empirical approach is taken where the Bayesian prior PDFs are generated from fits to the light-curve distributions themselves. In addition to the wavelet-based MAP approach to systematic error removal, PDC applies several other corrections to the data as described herein. This chapter is largely an updated version of citetstumpe:pdcArch.

### 8.1 The *Kepler* SOC Pre-Search Data Conditioning Pipeline Module

In order to search for planets, the *Kepler* Pipeline must first produce systematic error-corrected light curves. As described in the previous chapters in Part II (see Figure 8.1), the raw pixels are first calibrated at the pixel level by the Dynablock (DYN) and Calibration (CAL) modules, and then photometry and astrometry are extracted from the calibrated pixels by Photometric Analysis (PA) compo-

ment. The final step to produce light curves is performed in the Presearch Data Conditioning module (PDC), where signatures in the light curves that correlate with systematic error sources from the telescope and spacecraft, such as pointing drift, focus changes, and thermal transients are removed. Additionally, PDC identifies and removes the Sudden Pixel Sensitivity Dropouts (SPSD) that result in abrupt drops in pixel flux with short recovery periods. This is often preceded immediately by a cosmic or energetic solar ray event, and is sometimes followed by an exponential recovery over the course of a few hours, but usually not to the same flux level as before. Such step discontinuities are identified separately from those due to operational activities, such as safe modes and pointing tweaks, and are mended using a sophisticated method described in a companion paper (Kolodziejczak & Morris, 2012). PDC also identifies residual isolated outliers and fills data gaps (such as during intra-quarter downlinks) so that the data for each quarterly segment are contiguous when presented to later pipeline modules. Finally, PDC adjusts the light curves to account for excess flux in the optimal apertures due to starfield crowding as well as the fraction of the target star flux in the aperture to make apparent transit depths uniform from quarter to quarter, as the stars move from detector to detector with each roll maneuver.



**Figure 8.1** PDC in the context of the architecture of the SOC. PDC identifies and removes time-varying instrumental signatures appearing in the light curves produced by the Photometric Analysis (PA) module prior to the transiting planet search (TPS). Additionally, PDC identifies and removes sudden pixel sensitivity dropouts (SPSD) and isolated outliers. PDC also corrects the flux to account for the fact that each photometric aperture fails to capture all the light from its target star, and to account for contamination due to other stars' point spread functions leaking into the photometric aperture.

## 8.2 Introduction

With the area ratio of the Earth relative to the Sun being  $\sim 1 \times 10^{-4}$ , detection of transits of Earth-like planets requires an unprecedented photometric precision of 20 ppm (Koch et al., 2010; Jenkins et al., 2010a). The search for the transits is further complicated by the presence of systematic errors in the photometric data, which occlude the transit signatures as well as other astrophysical

signals in the light flux series. The main causes for the systematic errors are changes in the focus of the photometer (caused by intrinsic and extrinsic thermal variations), differential velocity aberration, residual spacecraft pointing errors<sup>1</sup>, mechanical vibrations, and electrical interference with other devices onboard the spacecraft (typically leading to moiré like patterns). Here we report on the new Presearch Data Conditioning (PDC) module in the *Kepler* Science Operations Center (SOC) Science Processing Pipeline, which is specifically designed to remove temporally correlated systematic errors and other artifacts from the light curves prior to planet detection.<sup>2</sup> The new PDC module, which has been completely rewritten for *Kepler* SOC 8.0 (the official *Kepler* software release in August 2011), shows dramatically improved performance – in particular at removing systematic trends while preserving stellar variability, and at detecting and correcting flux discontinuities.

In this chapter we describe the architecture and algorithms of the Presearch Data Conditioning module in the final software release, SOC 9.3, used to generate the legacy archival systematic error-corrected *Kepler* light curves. PDC has been completely rewritten and replaces the previous SOC 7.0 PDC module documented in Twicken et al. (2010b). We will briefly describe the different kinds of systematic errors typically encountered in the light curves, compare the main systematic error cotrending<sup>3</sup> algorithm of the current version with that of the old version, explain the software architecture of the PDC module, and show real light curve examples of the vastly improved performance of the new PDC module. Since one of the major improvements of this version of PDC is the Bayesian maximum *a posteriori* (MAP) approach to cotrending, we will also refer to the new version PDC first introduced in *Kepler* SOC 8.0 as “PDC-MAP”, and to PDC pre-8.0 as “PDC-LS” (where LS stands for “least squares”), when comparing the two<sup>4</sup>. Note that this chapter is largely based on Stumpe et al. (2012), Smith et al. (2012), and Stumpe et al. (2014), with updates and revisions reflecting the final, as-built software.

### 8.2.1 Errors in the Light Curves: Tasks of PDC

*Kepler*’s light curves as they come from PA contain a variety of systematic and other errors, which can occlude the astrophysical features in the light curves, and prevent detection of small planet transits. These errors are caused by a variety of instrumental effects and span a wide range of frequencies and amplitudes.

<sup>1</sup>*Kepler*’s pointing stability is  $0.009''$ ,  $3\sigma$  on timescales of 15 min and greater.

<sup>2</sup>PDC does not perform well for non-temporally correlated systematics that are unique to small numbers of stars such as moiré pattern noise.

<sup>3</sup>In co-trending, time series representing systematic or instrumental signatures are projected out of the light curves algebraically. This operation is distinct from detrending where low-frequency signatures are removed by an arbitrary filter designed to retain high frequency content.

<sup>4</sup>Note however that PDC-MAP and PDC-LS refer to the whole PDC module, rather than only the MAP or LS fitting algorithms.

The highest amplitude errors ( $\sim 1\%$ ) result from a combination of Differential Velocity Aberration (DVA) and long term focus changes, which can introduce low-frequency systematic errors over the whole quarter as the orientation of the Sun with respect to the telescope varies as *Kepler* orbits the Sun. Another source of strong systematic errors, typically with periods of hours to days, are focus changes that are caused by thermal transients. Such temperature changes are usually observed during the recovery from planned or unplanned interruptions in the science data collection. Unplanned interruptions include safe mode events and loss of fine point. Planned events that interrupt science data collection are the monthly “Earth-points” and the quarterly rolls during which *Kepler* points its high gain antenna (HGA) toward the Earth to downlink the previous month of data. In addition, the spacecraft is rotated  $90^\circ$  about its boresight during the quarterly breaks to keep the solar arrays and sunshade properly aligned towards the Sun and the radiator that cools the focal plane electronics pointed away from the Sun (Haas et al., 2010). Earth-points last about one day whereas quarterly rolls can take up to a few days if the spacecraft status is nominal. Temperature-change related focus changes can also be faster, such as the intermittent modulation of the focus by  $\sim 1 \mu\text{m}$  every 3.2 hours by a heater on one of the reaction wheel housings, which was partially mitigated after quarter 1 (Q1 – Jenkins et al., 2012). The magnitude of temperature-related effects on the focus has been found to be  $\sim 2.2 \mu\text{m}$  per  $^\circ\text{C}$  (Jenkins et al., 2012). Fast modes in the systematic errors can also be caused by thermal transients resulting from the desaturation of the reaction wheels on a three-day period and by pointing inaccuracies during the zero-crossings of the reaction wheels (due to enhanced rumble during these episodes).

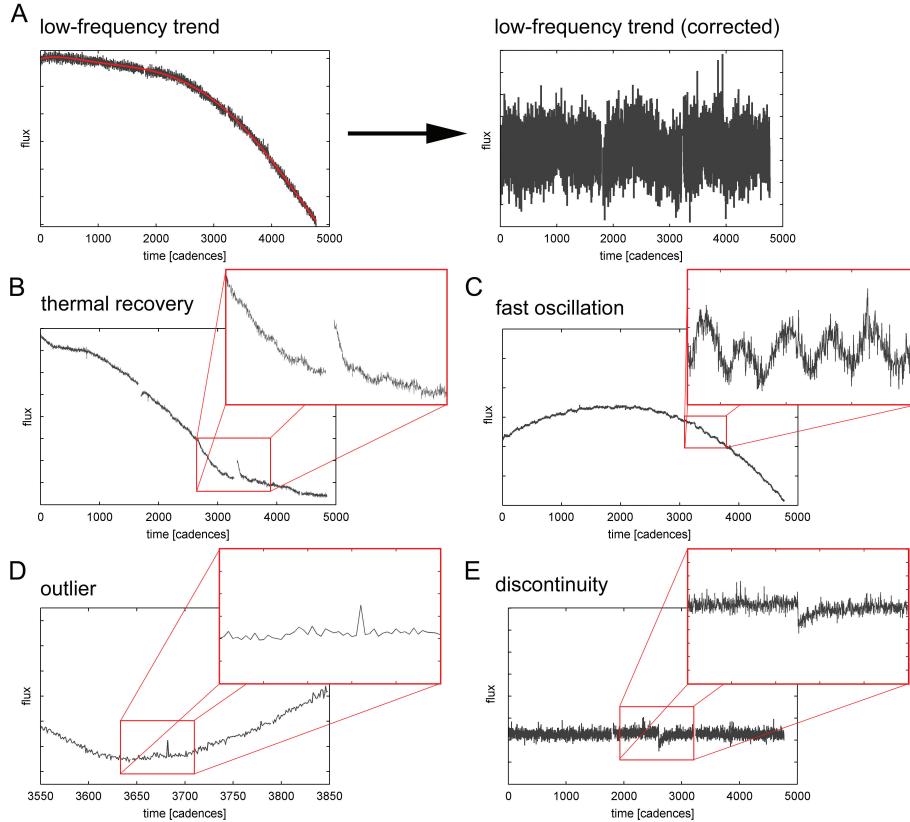
In addition to these systematic trends and modulations, the PA light curves suffer from local errors such as outliers, discontinuities, Argabrightenings (Witteborn et al., 2011), gaps in the data, and electronic image artifacts such as moiré pattern noise and rolling bands (see Chapter 4). Further, there are also global corrections that have to be applied to the flux amplitude, such as corrections for incompletely captured target flux (the flux fraction, see Subsubsection XIII of Subsection 8.3.3) and excess light from neighboring stars into the aperture (crowding metric). These effects are illustrated in Figure 8.2 and described in more detail below.

The task of the PDC module in *Kepler* is to identify and correct all these different kinds of systematic errors, while preserving planet-transits and other astrophysical signals in the light curves.

## 8.3 Architecture and Algorithms of PDC

### 8.3.1 Cotrending of Systematic Errors in PDC

The main component of PDC-MAP for removal of systematic trends in the light curves, and arguably the most determinant factor for the significantly improved



**Figure 8.2** Examples for typical errors in PA light curves, which are corrected in PDC. **A:** long trend (commonly due to DVA). **B:** thermal recovery, primarily an exponential recovery after an Earth-point, quarterly roll, or safe mode. The corresponding gap of approx. 50 cadences can be seen right before the recovery. **C:** fast oscillations ( $\sim$ 3-day cycle), such as from the reaction-wheel desaturation. **D:** outlier – possibly a cosmic ray. **E:** discontinuity – a sudden pixel sensitivity dropout (SPSD), commonly caused by radiation damage via cosmic rays or solar energetic particles. From Figure 2 of Stumpe et al. (2012).

correction performance as compared to PDC-LS, is the new cotrending routine based on a Bayesian maximum *a posteriori* approach (DeGroot, 1970). In general, removal of systematic trends in *Kepler* data is based on the assumption that the systematic error component in a light curve has a significant degree of correlation with a set of other time series that can be exploited to identify and remove the systematic error. For PDC-LS, this set of other time series was given by instrument readings on the *Kepler* spacecraft, whereas for PDC-MAP correlations with other light curves are used. We will give a brief review of the PDC-LS cotrending algorithm before describing the new PDC-MAP algorithm introduced in *Kepler* SOC 8.0.

**8.3.1.1 Cotrending in PDC-LS: Least squares fitting to ancillary engineering data – Stellar Variability lost** Removal of systematic errors in PDC-LS was previously performed based on correlations with a set of ancillary engineering data. These data include the temperatures at the local detector electronics below the CCD array

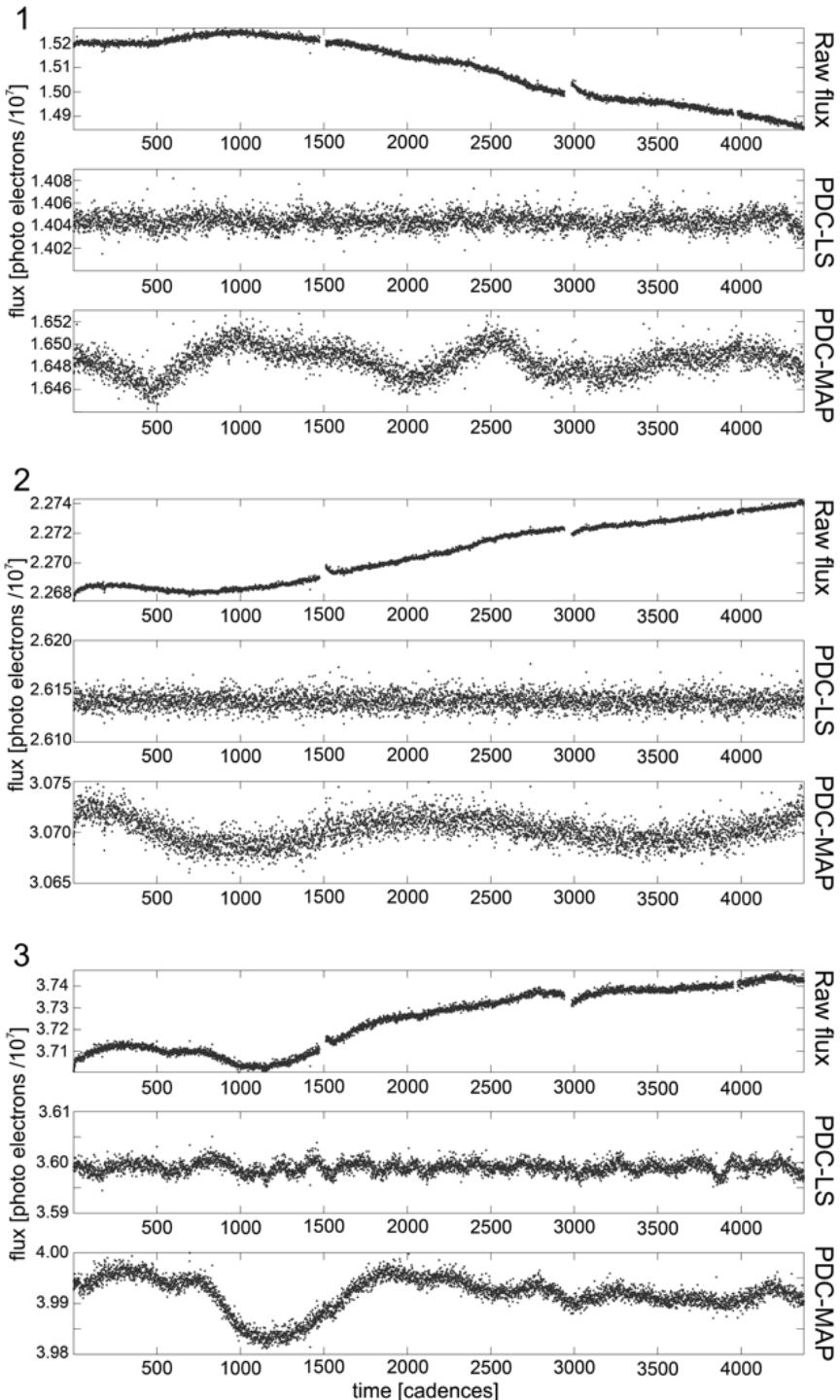
and polynomials describing the centroid motion of the targets from PA (Twicken et al., 2010b). This approach is based on the assumption that systematic errors in the light curves can be explained by a combination of instrument effects that are expressed in the engineering data and the motion of target centroids. A robust least squares fit of each light curve to the design matrix constructed from these time series is performed to identify and remove correlated signatures. For further details of the algorithm, see Twicken et al. (2010a).

PDC-LS did a good job at removing systematics in the light curves and allowed for many of the plethora of new discoveries in planet detection (Borucki et al., 2010; Holman et al., 2010; Doyle et al., 2011; Fressin et al., 2012; Lissauer et al., 2011; Welsh et al., 2012; Borucki et al., 2012) and stellar astrophysics (Meibom et al., 2011; Beck et al., 2011; Stello et al., 2011; Chaplin et al., 2011) based on *Kepler* data. However, it suffered from two deficiencies for a considerable fraction of light curves.

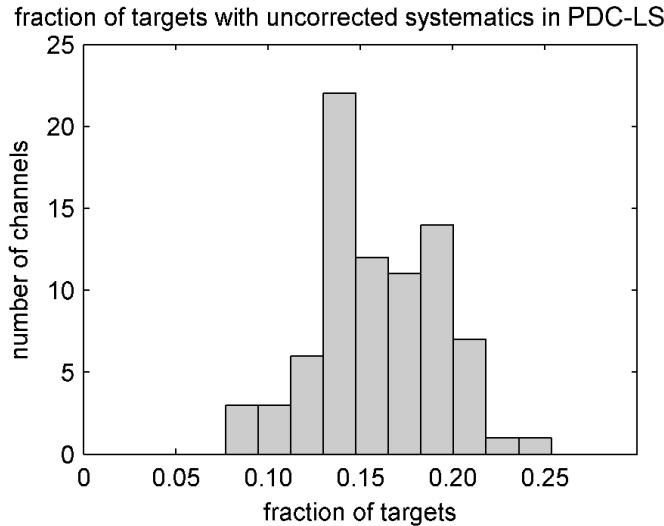
First, the biggest problem observed with PDC-LS was overfitting of the data, which led to removal of stellar variability – the corrected light curves of many stars appeared overly flat and featureless. This overfitting was due to the least squares approach to cotrending the light curves. With sufficiently many cotrending basis vectors<sup>5</sup>, coincidental correlations of instrument readings with stellar variability happened frequently, and thus the stellar features in the light curves were mistakenly identified as systematic errors and consequently removed. The Bayesian maximum *a posteriori* approach in PDC-MAP is specifically designed to prevent overfitting and in fact solves this problem very well, as will be shown below. See Figure 8.3 for some examples of the input light curve to PDC (the uncorrected flux output from PA), the (overfitted) output of PDC-LS, and the (significantly improved) output of PDC-MAP. Note that the PDC-LS and PDC-MAP light curves differ slightly in absolute magnitude, because the flux fraction correction (see Subsubsection XIII) was performed differently in PDC-LS.

A second problem of PDC-LS was that high-frequency noise was sometimes injected by the correction as an unwanted side effect of reducing the bulk root-mean-square (rms) deviation. A self-check was in place to prevent too strong noise-injection, in which case PDC-LS rejected the systematic error correction and instead returned a light curve where systematic trends were not removed, but only other corrections (e.g. discontinuities, outliers) were applied. The fraction of targets not cotrended at all was considerable. For instance, of all 162,926 targets observed with *Kepler* during Q7, 25,411 (15.6%) light curves were not cotrended. The distribution of non-cotrended targets per channel is displayed in Figure 8.4 and ranges from 7.7% to 25.4% per channel. Figure 8.5 shows some examples of light curves that could not be corrected properly with PDC-LS. The problem of noise-injection has been significantly reduced in PDC-MAP, as can be seen in the bottom panels of these examples.

<sup>5</sup>The cotrending basis vectors are a set of time series that describe the systematic errors and are used to remove these systematic errors from the light curves by fitting the light curves against the basis vectors.



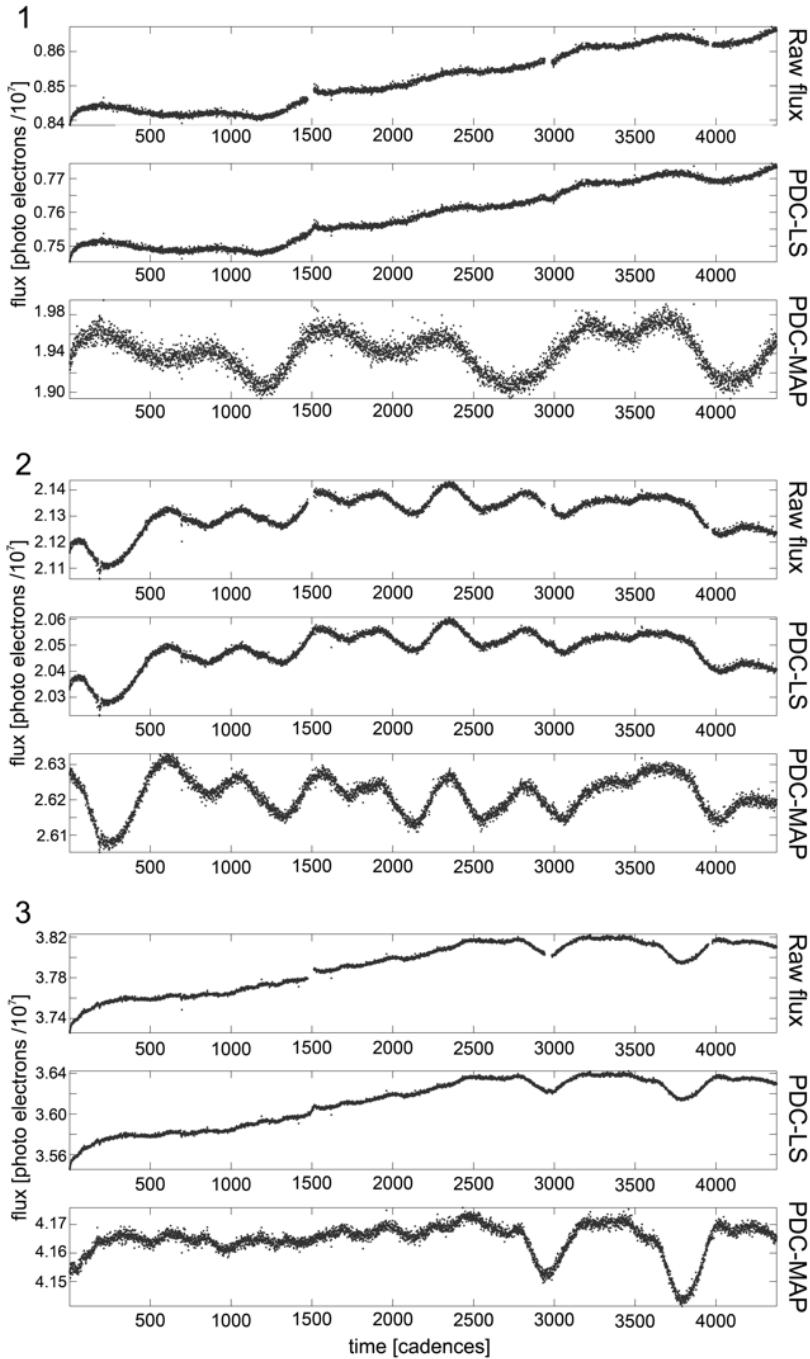
**Figure 8.3** Examples of light curves where the original PDC-LS overfitted the systematic trends and removed stellar variability. In each of the three panels (1–3), top is the input to PDC (output from PA), middle is PDC-LS (pre-Kepler SOC 8.0) using least squares cotrending, bottom is PDC-MAP using maximum *a posteriori* cotrending. Example 3 also shows incomplete cotrending of smaller scale feature in PDC-LS, as can be seen on the residual 3-day reaction wheel cycle signature (see Figure 8.2C). From Figure 3 of Stumpe et al. (2012).



**Figure 8.4** Fraction of targets per channel for which PDC-LS with least squares fitting to ancillary engineering data failed (Q7 data).

**8.3.1.2 Cotrending in PDC-MAP: Bayesian maximum *a posteriori* fitting to the ensemble of light curves – Stellar Variability regained** PDC-MAP uses a different approach to cotrending out instrumental signatures than was used in PDC-LS in mainly two regards. First, instead of employing ancillary engineering data, cotrending is performed against the ensemble of light curves on the same channel. Secondly, rather than using a simple least squares fit, a Bayesian maximum *a posteriori* (MAP) approach is used for fitting. Note that while this approach does not use any engineering data explicitly, it is still based on the assumption that instrumental signatures are the main cause for systematic errors in the data. Moreover, it assumes that these signatures are highly correlated among targets that are in proximity on the same CCD channel. This second assumption is a critical ingredient for the Bayesian MAP fit, as it allows for the generation of constraints for the fit by using information about systematics in targets within a neighborhood of the target under investigation. It is mainly these constraints that prevent overfitting, and thus help to preserve stellar variability. Here we will only provide a brief overview to an extent that is relevant in the context of the rest of this section.

Cotrending in PDC-MAP is performed by fitting each light curve to a set of basis vectors, and constraining the fit coefficients by using prior information about the probability density functions (PDF) of the coefficients. The basis vectors are generated from the 50% most correlated target light curves on the channel in order to have only the strongest correlated trends in the basis vectors and to largely exclude targets with substantial individual fluctuations due to stellar variability. Using Singular Value Decomposition (SVD) on this set of light curves, the  $N$  largest singular vectors are used as basis vectors. The number of basis vectors



**Figure 8.5** Examples of light curves where PDC-LS did not perform cotrending because too much noise was injected in the correction. Instead, the uncorrected flux from PA was returned, with only non-trend corrections performed, such as discontinuities, outliers, and flux magnitude. From Figure 4 of Stumpe et al. (2012).

used can be specified as an input parameter to PDC, with eight basis vectors being the default. By construction, these  $N$  basis vectors represent the majority of

the systematic errors contained in the light curves. Each of the light curves is projected onto these  $N$  basis vectors using a least squares fit. To make this fit more robust and prevent overfitting; however, the fit is constrained using prior information from targets expected to be affected by similar systematic errors. For that purpose, a target neighbor list is generated for each target with respect to their distance in a 3-D space spanned by stellar magnitude ( $Kp$ ), Right Ascension (RA), and Declination (Dec). The latter two coordinates, RA and Dec, directly translate into the spatial proximity of two targets on the CCD. Further,  $Kp$  is used as an additional dimension because systematic errors are not only correlated with location on the CCD but also with signal amplitude.

The contribution of each basis vector  $j$  to each light curve  $i$  is described in the fit coefficient  $\Theta_j^i$ . However, instead of simply using the fit coefficient of the least squares fit, the most likely coefficient given the observations in the neighborhood of the target is picked. In Bayesian terminology, the posterior PDF,  $p(\Theta|y)$ , is maximized, given a prior PDF  $p(\Theta)$  generated from fit coefficients of all neighboring targets, weighted by their distance to target  $i$ :

$$p(\Theta|y) = \frac{p(y|\Theta) \cdot p(\Theta)}{p(y)} = \frac{p(y|\Theta) \cdot p(\Theta)}{\int p(y|\Theta)p(\Theta)d\Theta}. \quad (8.1)$$

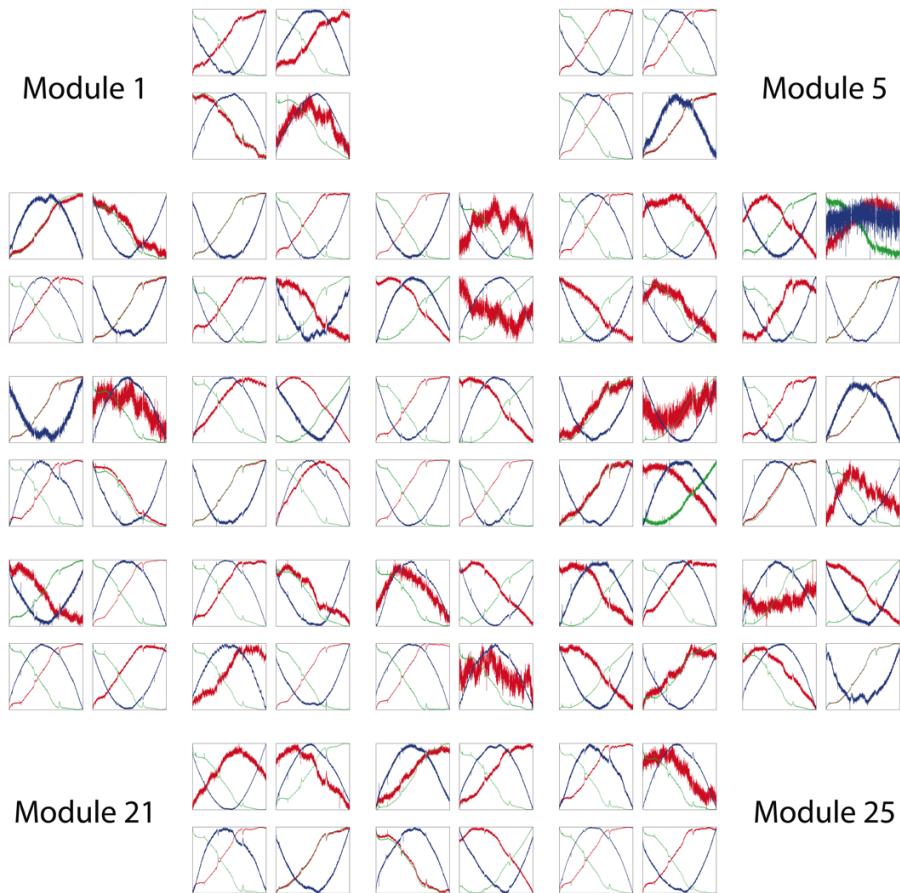
The  $\Theta$  that maximizes  $p(\Theta|y)$  is used as the fit coefficient for the MAP fit. Note that the denominator  $p(y)$  is simply a normalization over all possible observations  $y$  and can in practice be omitted in the maximization process. Taking the maximum *a posteriori*  $\Theta$  effectively constrains the fit and helps to avoid overfitting. This is a form of regularization or shrinkage in other contexts. In particular, only light curve features also present in targets in the neighborhood will be removed, whereas any stellar variability that might have coincidental correlation with some of the basis vectors will be constrained by the prior, and hence be preserved.

Figure 8.6 shows the first three basis vectors for each channel on the CCD for Q7. This illustrates how the dominant systematic errors vary across the field of view (FOV) of the CCD. However, some trends are observed on almost all channels.

### 8.3.2 Inputs to PDC

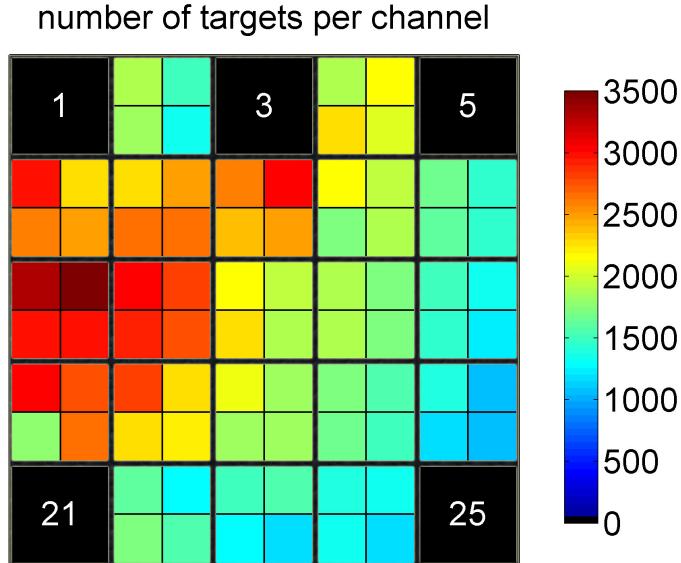
The inputs to PDC are the calibrated, uncorrected flux time series (measured in  $e^- s^{-1}$ ), as prepared by PA (see Chapter 6). One unit of work for PDC is a single module output (or readout channel) from the CCD with the duration of nominally three months for LC data, with a sampling period of 29.4 minutes. Thus, there are typically 1000–3500 targets per channel (see Figure 8.7), and about 4500 data points (“cadences”) per target<sup>6</sup>. In addition to the flux time series, the target data

<sup>6</sup>This is for long cadence (LC) data, which is addressed here. Short cadence (SC) data is sampled at a rate of 57.8 seconds for the duration of ∼one month and there are only 512 such targets in total on all 84 module outputs.



**Figure 8.6** First three cotrending basis vectors for each channel on the *Kepler* CCD, data from Q7. Note that the four channels of module 3 are missing because that module failed during Q4 in January 2010. From Figure 6 of Stumpe et al. (2012).

structure also contains per-cadence flux uncertainty values and gap indicators to denote invalid cadences, per-target data such as the flux fraction and crowding metric (see below), as well as bookkeeping information such as the *Kepler* ID (the “KIC” number) and other stellar parameters of interest.<sup>7</sup>



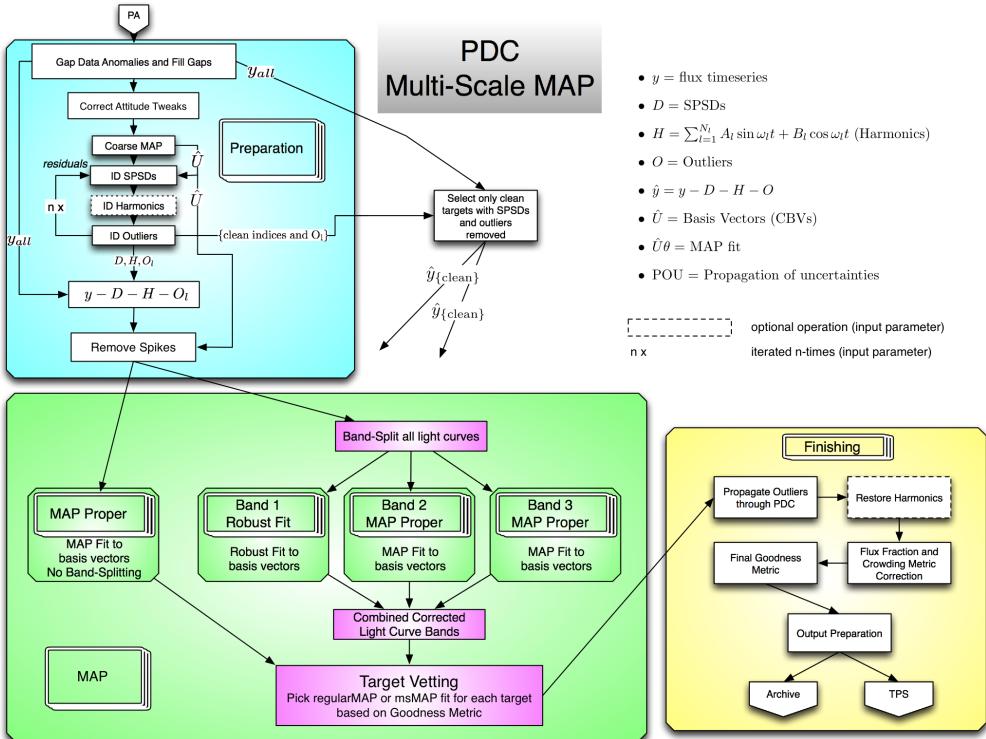
**Figure 8.7** Number of targets per channel in Q7. Each of the 25 modules is divided into four module outputs (channels). Modules 1, 5, 21, and 25 are used for fine pointing control and not for collection of science data. Module 3 failed in Q4 in January 2000, leaving a total of 80 channels. There are significantly more targets in the upper left corner because that section of the FOV is closer to the galactic plane during this season. From Figure 7 of Stumpe et al. (2012).

### 8.3.3 Overview and Data Flow

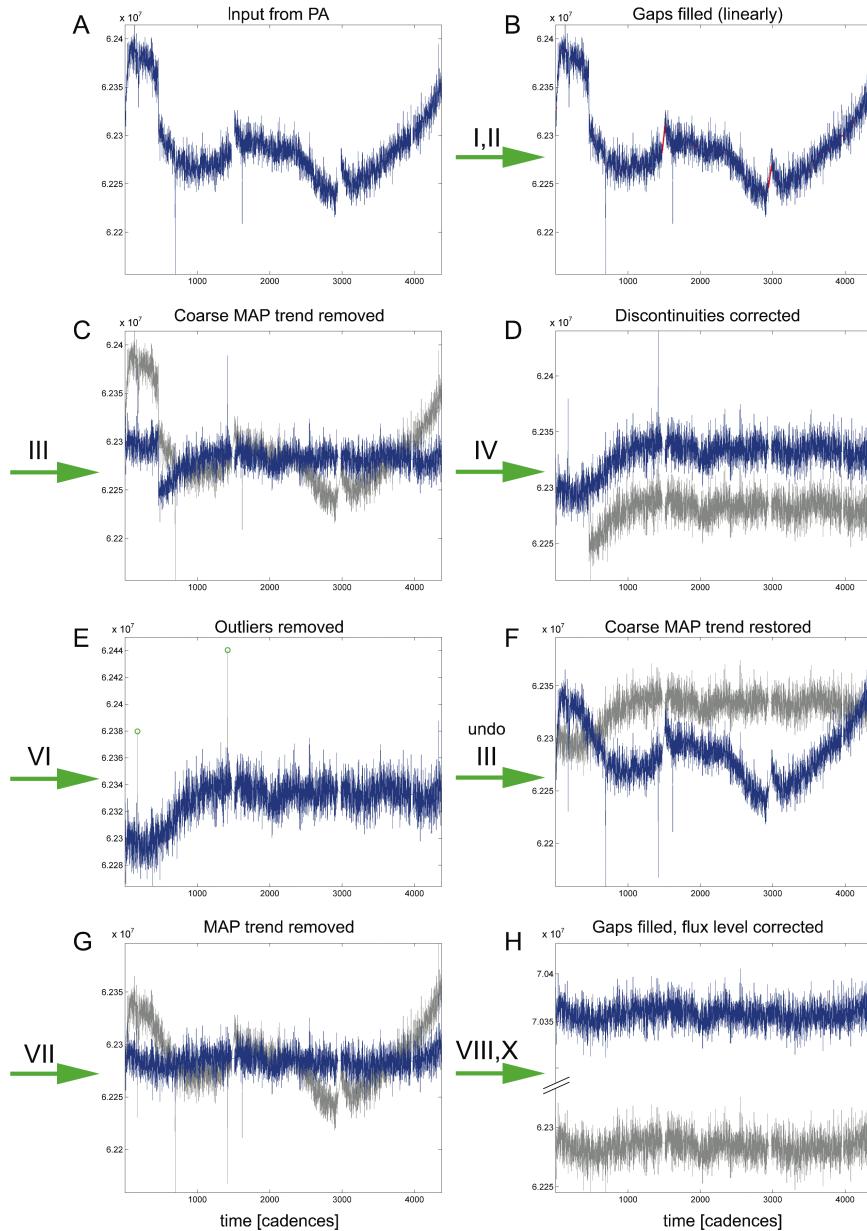
The data flow in PDC-MAP is displayed in Figure 8.8. Figure 8.9 and Figure 8.10 show two light curve examples as they are being processed through PDC. The individual operations of this sequential processing are described below:

**I Gap Data Anomalies** The first step is to flag anomalies in the flux series and mark the respective cadences as not usable (i.e., “gapped”). Anomalies may affect only one target or all targets on the channel (e.g., monthly downlinks, quarterly rolls, attitude adjustments, spacecraft safe modes, loss of fine-pointing, Argabrightenings (Witteborn et al., 2011), and can have a duration of one to several hundred cadences. The information about the gaps and the type of anomaly is provided separately by PA. Gapped data is labeled as such and is not exported to the Mikulski Archive for Space Telescopes (MAST).

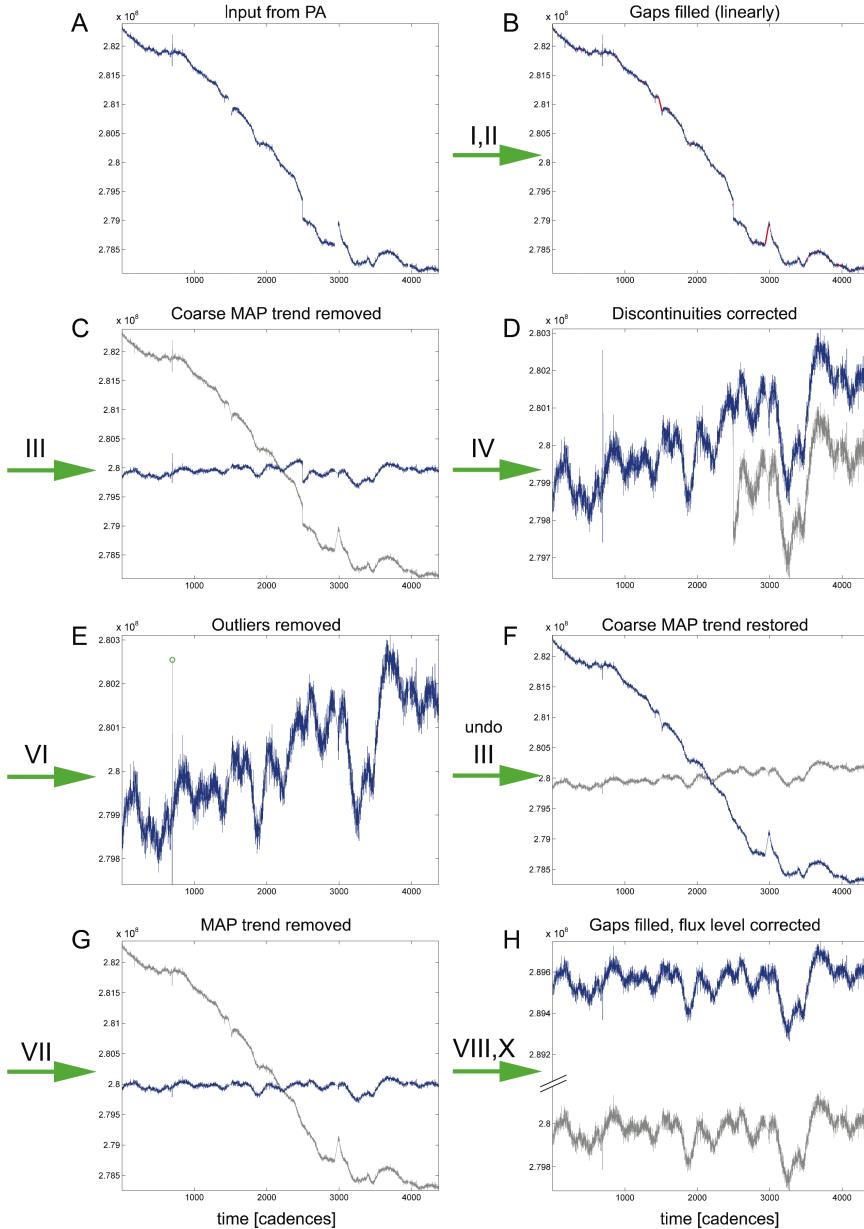
<sup>7</sup>The updated *Kepler* Input Catalog (KIC) is the primary source of information about targets observed with *Kepler*, and contains their RA/Dec coordinates, estimates for a star’s radius, temperature, surface gravity, and other information. URL: [http://archive.stsci.edu/kepler/kepler\\_fov/search.php](http://archive.stsci.edu/kepler/kepler_fov/search.php)



**Figure 8.8** Architecture and Data Flow of PDC-MAP as of *Kepler* SOC 8.0. From Figure 8 of Stumpe et al. (2012).



**Figure 8.9** Example for light curve processing of a quiet star in PDC from the output of PA and input to PDC (top right) to the output of PDC (bottom right). The numbers above the arrows denote the operations as defined in Subsection 8.3.3 to get from the previous state (which is shown again in gray for comparison) to the current state (blue curve). **A:** Input to PDC. **B:** Gaps have been linearly filled. **C:** The coarse MAP trend has been removed, a temporary correction to facilitate discontinuity correction and outlier removal. **D:** Discontinuities have been corrected. **E:** Outliers have been removed. **F:** The coarse MAP trend has been restored. **G:** The final MAP fit has been removed. **H:** Gaps have been filled using a more sophisticated algorithm and the absolute flux magnitude has been adjusted by flux fraction and crowding metric corrections. From Figure 9 of Stumpe et al. (2012).



**Figure 8.10** Example for light curve processing of a variable star in PDC from the output of PA and input to PDC (top right) to the output of PDC (bottom right). The numbers above the arrows denote the operations as defined in Subsection 8.3.3 to get from the previous state (which is shown again in gray for comparison) to the current state (blue curve). **A:** Input to PDC. **B:** Gaps have been linearly filled. **C:** The coarse MAP trend has been removed, a temporary correction to facilitate discontinuity correction and outlier removal. **D:** Discontinuities have been corrected. **E:** Outliers have been removed. **F:** The coarse MAP trend has been restored. **G:** The final MAP fit has been removed. **H:** Gaps have been filled using a more sophisticated algorithm, and the absolute flux magnitude has been adjusted by flux fraction and crowding metric corrections. From Figure 10 of Stumpe et al. (2012).

**II Fill Gaps** Even though gapped data is not being exported, some operations in PDC require contiguous data over the whole time series. Therefore, gaps are filled by piecewise cubic Hermite interpolation (PCHIP) for internal use. Since PCHIP does not work well at extrapolation, if there happens to be any gapped cadences at the beginning or end of the time series, a linear nearest-neighbor approach is used to fill those cadences separately. Note that these simply interpolated gap values are filled a second time in TPS, using a more sophisticated algorithm to prepare the light curves for planet searching.

**III Correct Attitude Tweaks** Occasionally during Q2 (20 June 2009 – 16 September 2009) the spacecraft drifted too far off true pointing and an “attitude tweak” correction was made. These adjustments result in a sudden shift in the pointing as illustrated in Figure 8.39. Although rare, they result in a dramatic perturbation to the time series. If these steps are presented to MAP then the basis vector will reflect their existence and PDC-MAP will do its best to remove them on each target. However, this correction was found to be poor if performed simultaneously with the other corrections in PDC-MAP. We therefore correct for them right at the beginning of the PDC procedure. All attitude tweak locations are available from mission spacecraft operations and supplied to PDC as a data anomaly flag.

This attitude tweak correction method first tests each tweak location on each target to see if a tweak actually exists. A simple step function filter is used to detect the tweak. If the response to the step function is greater than  $4\sigma$  compared to the neighborhood about the tweak then a tweak is detected and the correction is performed. The reason for the existence test is because targets with fast oscillating stellar signals can look like steps at the tweak. The tweak correction would then try to remove both the actual tweak (if it exists) and the oscillation at the tweak. So, with the filter test, only tweaks that appear large compared to the oscillation signals about the tweak are attempted to be removed. This means for many highly oscillating targets the tweaks will not be removed (nor visible in the data).

Another potential issue is the transits of giant planets. If there are giant transits near the tweak then the filter detector will also falsely *not* detect the legitimate tweak. To protect from this behavior all known cadences within transits are filled using the same gap filler as in Subsubsection II above before the step filter is applied.

For those tweaks that pass the filter, the following three steps are used to fix the attitude tweak:

1. Apply a Savitsky-Golay filter to smooth the data before and after the tweak (separately, in two parts, so not to smooth out the tweak!). The flux value difference between the end of the first part before the tweak and the beginning of the second part after the tweak is used to find the offset due to the tweak. The smoothing is not saved to the flux, but is just intended to find the offset.

2. Add the found offset to all cadences after the tweak to line up the ends about the attitude tweak.
3. The attitude tweak cadence itself can have some intermediate flux value between the before and after values. We therefore interpolate over the single tweak cadence using simple linear interpolation.

**IV Coarse Cotrending** A first coarse cotrending is performed, using the Bayesian maximum *a posteriori* (MAP) approach described in Section 8.4. This coarse cotrending is only applied temporarily to the flux series to enhance the performance of the next three (Subsubsection V, Subsubsection VI, Subsubsection VII), and then the coarse correction is restored in Subsubsection IX afterwards in order to perform the proper MAP correction in Subsubsection XI.

**V Sudden Pixel Sensitivity Drop-out (SPSD) Correction** Approximately 3% of the light curves contain one or more noticeable downward step discontinuities per quarter. The vast majority of these discontinuities are the result of cosmic rays or solar energetic particles striking the photometer and causing permanent local changes in CCD pixel sensitivity, although partial exponential recovery is often observed (Jenkins et al., 2010b). Depending on the incident angle and the energy of the cosmic ray, up to a few pixels can be affected, and the net decrease in quantum efficiency is typically  $\sim 0.5\%$  for a light curve. Since these SPSDs are not correlated systematics, they can not be removed by the MAP algorithm (see Section 8.4), and instead a specific submodule exists to detect and correct SPSDs. The discontinuity correction algorithm used in PDC-LS had a considerable false-negative rate, and was moreover performing a rather crude correction that did not take partial recoveries into account. The new SPSD module in PDC-MAP is significantly more sophisticated and performs markedly better at detecting and correcting SPSDs. The details of this new algorithm are described in KADN-26304. In brief, SPSD detection is performed by fitting a window of the data to a combination of basis functions, with one of the basis functions modeling a step discontinuity in the flux. The fit coefficient of this basis function is evaluated in the context of flux variability of the light curve ensemble on that channel, to determine the probability for an underlying SPSD. Further checks are performed to prevent false positive detection of transit-like features as discontinuities. Correction then proceeds as a two-stage process, with the first stage estimating the step size (based on analysis of the entire flux series) and the second stage modeling the recovery process. The sum of these two components is the total correction applied to the light curve. In the Q9 data, SPSDs were detected on 5,252 out of 167,404 targets (3.1%). Figure 8.11 shows a histogram of the fraction of targets with detected SPSDs.

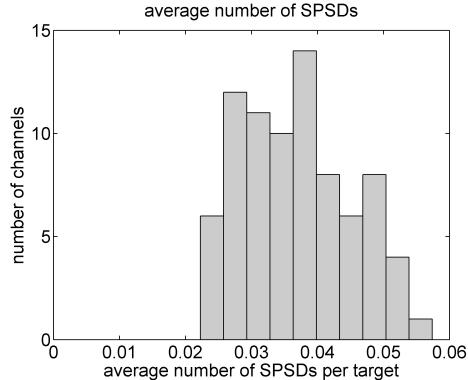
**VI Harmonics Removal (optional)** Harmonic content can be identified and removed from the light curves of harmonically variable targets. This step was a mandatory operation in PDC-LS in order to prevent harmonic content from corrupting

the least squares fitting in the cotrending routine. In PDC-MAP the maximum *a posteriori* fitting approach is robust against uncorrelated target-specific harmonics and this step is therefore not performed per default but can be requested with an input parameter. If harmonics are removed for further processing in PDC they are restored later after cotrending.

From a theoretical point of view, one might expect that this (previously required) operation of harmonics removal prior to cotrending would improve the cotrending – simply because it reduces the feature complexity of the light curves. However, we found that it has virtually no effect with the new MAP cotrending method. In a direct comparison of 1075 light curves processed with and without harmonics removal, we found no difference at all for 97% of the targets. For the remaining 3%, performing harmonics removal before cotrending resulted in weak residual long-term trends, which had been identified as very low-frequency harmonics by the harmonics detector. Thus, by removing them before cotrending, and restoring them afterwards, these long-term trends were preserved in the light curves. Without knowing the ground truth, one obviously cannot know with certainty whether these long-term trends are actually astrophysical effects (in which case they should be preserved) or systematics errors (in which case they should be removed). However, for most of these cases the low-frequency trends seemed to be systematic errors, as judged by comparison with the systematic trends identified for other targets, and therefore performing harmonics removal appeared to actually decrease the quality of the error correction.

In summary, while harmonics removal was a required step in the least square based cotrending in PDC-LS, it seems that this operation does not improve the error correction (and may even do harm in a few cases) in the new Bayesian approach of PDC-MAP. Therefore, this step is skipped by default.

**VII Outlier Correction** Outliers in each flux time series are identified by thresholding a sliding window-based detrended time series. The time series is detrended by splitting it into small chunks of 72 hours in length and robustly fitting polynomials to each chunk. The optimal polynomial order for a fit to each chunk is determined using the AI criteria. This fitting is then repeated after shifting the chunks by half a chunk. The fitted time series for the shifted and non-shifted cases are then added together after tapering using a Bartlett window. This ro-



**Figure 8.11** Fraction of targets per channel for which an SPSD has been detected in Q9. The total number of channels is 80, and not 84, because module 3 failed in Q4, reducing the number of data channels by four.

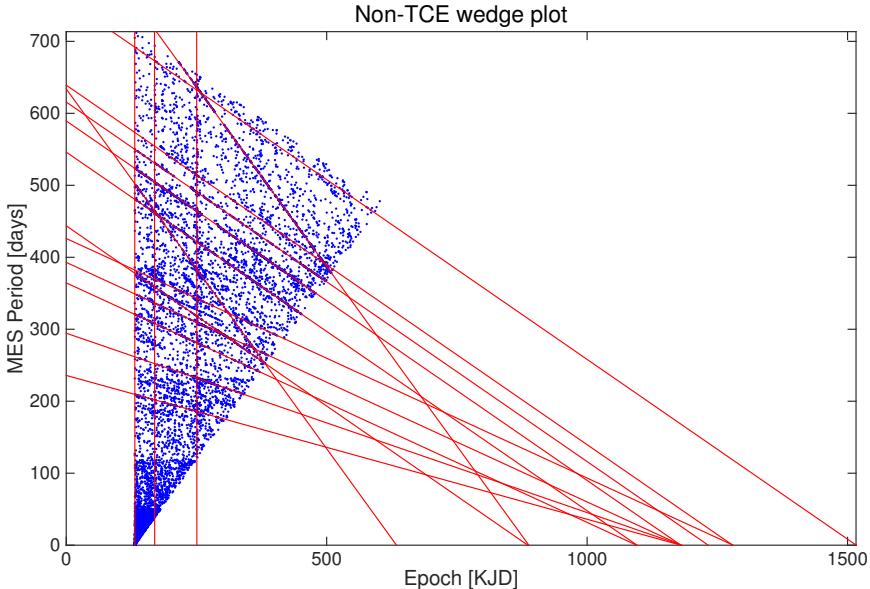
bust fit polynomial time series is then subtracted from the original time series to get the residuals. All cadences 10 times the median absolute deviation of the residual are then flagged as outliers. Only isolated (single cadence) points above the threshold are considered to be outliers in order to prevent flagging of planet transits or stellar flares as outliers. These outlier values are corrected by replacing them with PCHIP interpolated values just as with the gaps in Subsubsection II. The purpose of this correction is to prevent false triggering of Threshold-Crossing-Events (TCE) in TPS and also to improve the performance of other operations downstream in PDC. Before export to the MAST, the original outlier values and uncertainties are restored to the corrected light curves and propagated through the PDC correction steps like all regular data points.

**VIII Iteration of SPSD Correction, Harmonics Removal, Outlier Correction** The operations of SPSD Correction (Subsubsection V), Harmonics Removal (Subsubsection VI), and Outlier Correction (Subsubsection VII) are non-orthogonal to each other. In theory, performing them simultaneously rather than sequentially should yield optimal correction, however at the expense of a substantially more complex algorithm, and so was not developed. Iteration of these operations is an approximation to such a joint fitting approach, and an input parameter was included to iterate the Subsubsection V–Subsubsection VII for improved correction performance. In particular, at the time of the design of the architecture it was expected that Harmonics Removal would improve the performance of the SPSD correction (Kolodziejczak & Morris, 2012) (Subsubsection V), and hence that iterating Subsubsection V–Subsubsection VII would be beneficial for the overall correction. However, it was found that the presence of harmonics in the frequency- and amplitude-range that is observed in the *Kepler* data does not decrease the detection performance of the SPSD module by more than 1%, while removing them would come with the disadvantages mentioned above for Subsubsection XI. For these reasons, the Subsubsection V–Subsubsection VII are per default only performed once in PDC-MAP, while leaving the option to perform an iteration in case future design changes should make this desirable.

**IX Restore the Coarse MAP Correction** As with the other corrections above, the ideal PDC method would apply all correction at the same time in a joint fit. As a first approximation to this we iterate MAP with the other corrections by first applying a coarse MAP correction before removing SPSD, harmonics, and outliers. We then restore the coarse MAP correction before proceeding.

**X Remove Systematic Spikes** The Transiting Planet Search (TPS – see Chapter 9) component has in the past identified many features in the light curves that consistently produce high SNR multiple event statistics (MES). TPS fortunately will veto most of these events; however, their presence results in the masking of potentially legitimate TCEs. It is therefore to our advantage to remove these systematic features before the transit search. An illustration of these systematic events is shown in the “wedge plot” in Figure 8.12. The figure plots the TPS

maximum MES trigger epoch (in KJD) versus period (in days) for 9650 random non-TCE targets. A Hough Transform is used to identify strong lines in the scatter plot. Each line is indicative of a single event in time that is causing systematic features on multiple targets. The x-axis intersection of each line indicates the epoch of the feature. These features are very short (1–3 cadence) impulses in the light curves. Such features are very difficult to remove using PDC-MAP, even in band 3. A completely separate method was therefore developed to remove these specific features.



**Figure 8.12** A “Wedge Plot” of 9650 random target TPS trigger showing the strong lines caused by systematic sharp features in the data.

The method first finds “spike basis vectors” derived from the coarse MAP basis vectors. The coarse MAP basis vectors are passed through a wavelet-based de-noiser to isolate strong features in the vectors using the following steps:

1. Extend the basis vectors by reflection to mitigate boundary effects.
2. Compute an overcomplete wavelet transformation of the basis vectors, which yields a set of wavelet coefficients at all relevant scales.
3. Apply a “universal hard threshold” (Donoho & Johnstone, 1994) to the wavelet coefficients at each scale. Coefficients with values below the threshold are presumed to be noise, and are therefore replaced with zeros.
4. Apply the inverse overcomplete wavelet transformation to the ‘thresholded’ wavelet coefficients to estimate the de-noised basis vectors.

The de-noised basis vectors are then reanalyzed to find the significant number of basis vectors using the method described in Subsection 8.5.1 (i.e. find the num-

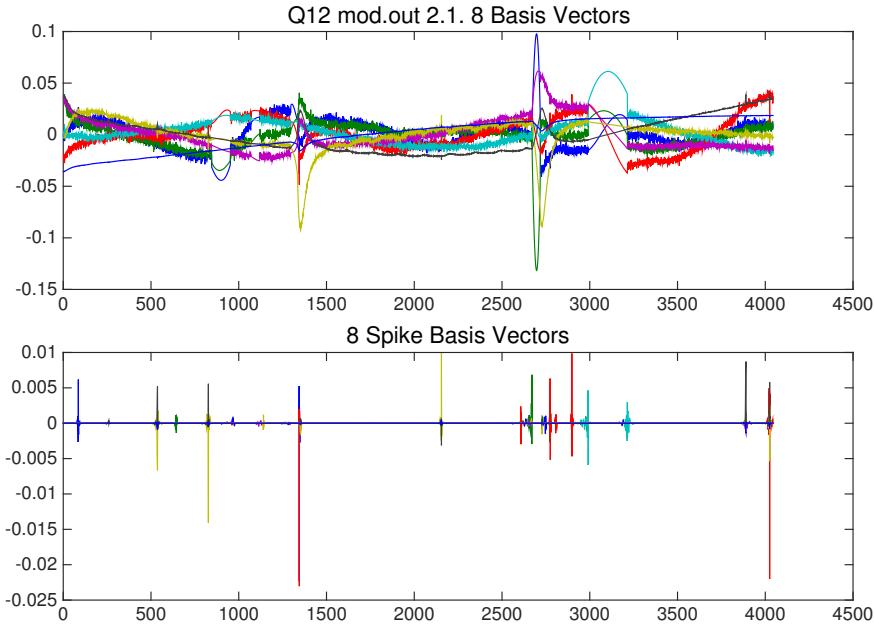
ber of basis vectors that still have signal after de-noising). Extracting the spike basis vectors from these de-noised coarse basis vectors involves the following steps:

1. Gap 10 cadences around each Earth-point so that they are not identified as spikes.
2. Pass the de-noised basis vectors through a 15-cadence median filter to isolate only high frequency signals.
3. Take the second derivative of the basis vectors (second differences) to identify sharp spikes.
4. Calculate the median absolute deviation (scaled by 1.254 to convert to standard deviation units) and flag all cadences above  $100\sigma$ .
5. Also flag all cadences within one cadence of an existing cadence flag in the previous step.
6. The spike basis vectors are the median-filtered de-noised basis vectors where all non-flagged cadences are zeroed.

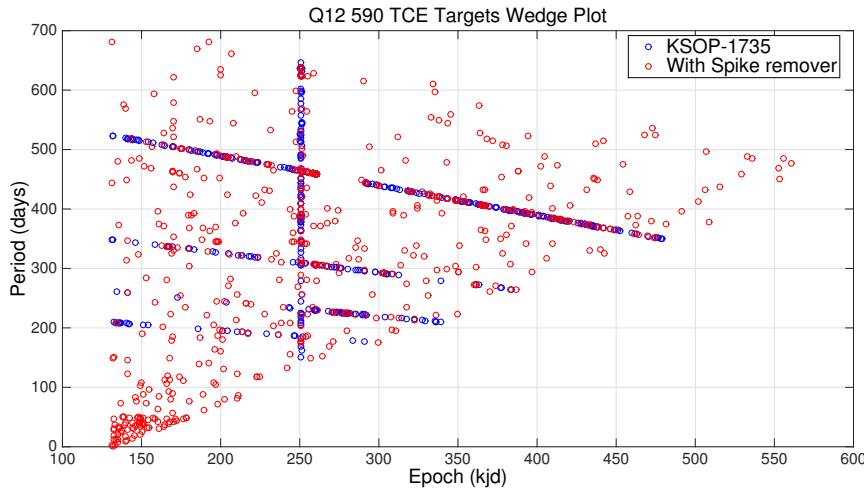
Note that the coarse MAP basis vectors are *not* altered in any way. Instead after the restoration of the coarse MAP correction we remove the spikes in the data separately. For each target we high pass filter the flux time series using the same 15 cadence median filter. This is to reduce the risk of overfitting the spike basis vectors. We then robust fit the spike basis vectors to the median filtered flux time series and then use the resulting robust fit coefficients to subtract the spike basis vectors from the flux time series after Subsubsection IX.

Figure 8.13 shows the coarse MAP basis vectors and the derived spike basis vectors. The spikes are mostly visible in the coarse basis vectors but are small compared to most features and therefore MAP does not effectively remove them. In isolation we can very effectively remove the spikes. An illustration of performance can be found in Figure 8.14. Here we plot in blue circles the wedge plot points for 590 TCE targets that all line up on lines. In red circles we plot the same targets but after applying the spike remover. We clearly see the targets have a much more evenly spread out distribution. About 40% of targets still lie on lines but the lesser extent illustrates that the systematic spikes have been greatly attenuated. The other 60% of the targets were “freed up” to search for other potential planet candidates.

**XI Bayesian MAP Approach for Cotrending to Remove Systematic Errors** This operation constitutes the main step for the correction of systematic trends in the light curves. Each flux time series is cotrended against the ensemble of light curves in order to remove correlated trends using a Bayesian maximum *a posteriori* approach as described in Section 8.4. The main cotrending procedure (this step) is the same as the routine performed in the Coarse Cotrending (Subsubsection IV),



**Figure 8.13** Q12 mod.out 2.1 Coarse MAP basis vectors and the derived spike basis vectors.



**Figure 8.14** Wedge Plot comparison of 590 TCE targets with and without the spike remover.

but with three important differences that make its correction significantly more accurate. First, only “clean” targets, for which no discontinuities had been found (Subsubsection V), are considered to generate the basis vectors. Secondly, discontinuities, harmonics (if enabled), and outliers have been removed from the light curves before the fit, which improves the quality of the correction. Thirdly, two applications of MAP are performed, a single-scale MAP as described in

Section 8.4 and a multi-scale wavelet-based band split MAP as described in Section 8.6.

Once finished we have two cotrended light curves 1) single-scale MAP and 2) multi-scale MAP. We select the best light curve for each target based on the goodness metric (Subsubsection XIV). We use three of the goodness metric components to make our selection: i) Introduced Noise, ii) Delta Variability, and iii) Earth-Point Removal. See Subsubsection 8.6.4.1 for details on the target vetting process.

**XII Restore Harmonics (optional)** If harmonic content had been removed from the light curves (optional Subsubsection VI), then it will be added back here, after properly cotrended with MAP. Per default, harmonics removal is not performed in PDC-MAP.

**XIII Crowding Metric and Flux Fraction Correction** When generating flux series from pixel data in PA, the optimal aperture (Twicken et al., 2010b) for each target may include additional light from nearby light sources. If not removed, this excess flux would decrease the apparent planet transit depth and lead to a systematic underestimation of planet radii. The so-called *crowding metric* for each target is computed in TAD and PA-COA, and reflects which fraction of the flux in the optimal aperture is due to the target itself. It is a scalar value representing the average over the effective date range. Note that using a scalar average for the crowding metric is an approximation since background sources of light may enter or exit the optimal aperture over the course of the three-month data period. Similarly, using a scalar value for the flux fraction is an approximation because the centroid of the point-spread-function (PSF) of a target can move over the course of a quarter<sup>8</sup>.

Similar to excess flux leaking into the optimal aperture, a fraction of the PSF of the target may not be captured in its optimal aperture. To account for this missing fraction, the *flux fraction*, computed by PA, is used to normalize the flux.

Together, these two corrections can be expressed as:

$$\tilde{y}(t) = \frac{y(t) - (1 - c) \cdot \text{median}\{y(t)\}}{f}, \quad (8.2)$$

where  $\tilde{y}(t)$  is the corrected cotrended flux,  $y(t)$  is the cotrended flux,  $c$  is the crowding metric, and  $f$  is the flux fraction. The *median* is used here, rather than the *mean*, to make the normalization robust against mathematical outliers. This is important in *Kepler* light curves because transits from planets or eclipsing binaries represent mathematical outliers.

**XIV Goodness Metric** Characterizing the performance of systematic trend removal in PDC can be highly subjective which emphasizes the difficulty in the

<sup>8</sup>The maximal motion of a target is  $\sim 0.6$  pixel over 3 months.

cotrending task. A human can peruse a selection of light curves and identify poorly cotrended examples to rate performance. Obvious trends and features can be evaluated but subtle issues with the performance will likely be overlooked. So, in an effort to minimize the subjective analysis, a numerical goodness metric has been developed. There are four critical aspects of the cotrending performance to consider in the metric: 1) removal of systematic trends, 2) no introduced noise, 3) preservation of stellar variability and 4) removal of the Earth-point thermal recovery. A successful cotrend occurs when all four conditions are met, so one cannot rely solely on target-to-target correlation, for example, to determine acceptable performance. Overfitting, and hence flattening of stellar features, results in zero correlation but would not be considered a “good” correction. We expect the goodness metric to represent a coarse estimate of the correction quality, rather than a perfectly accurate quantification of the correction performance (which in fact would require the ground-truth light curves). In particular, the goodness metric should help to achieve four goals:

1. For each light curve (e.g. processed with both regular MAP and msMAP), to decide which of two possible corrections is more likely to be the better correction.
2. For each module output decide which set of PDC parameters on average produces better results.
3. To detect cases where the correction of a target was very poor, so that the respective target can be corrected differently or investigated further.
4. To give users of *Kepler* data an estimate of the quality of the PDC correction for any particular light curve.

The goodness metric calculates the cotrending performance in each of the four components using the following methods developed experimentally to agree with observations of critical features.

Removal of systematic trends is evaluated by the target-to-target correlation. If the correlation is near zero then in principle no systematics can remain<sup>9</sup>. The Pearson correlation,  $C$ , is computed over the entire module output and the correlation goodness,  $G_c$ , is computed as the arithmetic mean of the cube of the absolute value of the correlation between the target under study and all other targets,

$$G_{C,i} = \alpha_C \frac{1}{N} \sum_{j \neq i} (|C_{ij}|^3) \quad , \quad (8.3)$$

where  $i$  refers to the target under study,  $j$  is summed over all other targets,  $N$  is the total number of targets, and  $\alpha_c$  is a scaling factor. The purpose of the cube is to over-emphasize any strong correlations.

<sup>9</sup>The root sum-square mean of off-diagonal correlations of white Gaussian noise is exactly  $1/\sqrt{N_{\text{sample}}}$  so even with no systematics the off-diagonal correlation matrix will not be exactly zero.

Introduced noise is determined by examining the change in the power spectral density (PSD) of the noise floor for each light curve before and after the correction:

$$\Delta \text{PSD}(freq) = \frac{\text{PSD}(\text{Nf}_{\text{after}})}{\text{PSD}(\text{Nf}_{\text{before}})}, \quad (8.4)$$

$$G_{N,i} = \alpha_N \int_{\Delta \text{PSD} > 1} \log(\Delta \text{PSD}(freq)) dfreq, \quad (8.5)$$

where Nf is the noise floor of the light curve, here defined as the first differences between adjacent flux values. The noise floor is used because removal of trends can result in large changes in the PSD, but signals close to the *Nyquist* frequency are dominated by noise. The integral is also only taken about regions where the change in power is greater than one, implying an increase in noise.

The third component, preservation of stellar variability, is the most difficult of the four to numerate. An exact determination would require knowledge of the intrinsic signal. Since it is not possible to make this determination, an estimate is made by assessing a mid-frequency band region where light curve signals are dominated by stellar features. High-frequency components are removed via a Savitzky-Golay Filter and the low-frequency components via a third-order polynomial removal. What is left is almost always overwhelmingly dominated by stellar signals.<sup>10</sup> The exception being the Earth-point recovery regions, which can be quite strong and so these regions are masked. The standard deviation of the difference between the resultant mid-frequency light curves before and after the correction,  $\tilde{y}_{\text{before}}$  and  $\tilde{y}_{\text{after}}$  respectively, is then computed:

$$G_{V,i} = \alpha_V \text{ std}(\tilde{y}_{i,\text{after}} - \tilde{y}_{i,\text{before}})^2 \sqrt{V_i}, \quad (8.6)$$

where  $V$  is the variability as calculated by PDC-MAP (see Equation 8.20) and is used to emphasize targets with greater variability.

The final component is the removal of the Earth-point thermal recoveries. Because the thermal transients after an Earth-point roughly follow an exponential shape, we base the calculation of this component on the strength of the exponential character of the light curve in this region. For that purpose, we pre-process the light curve by normalizing it with its median and performing a simple linear detrending. Then a non-linear least-squares fit is used to fit an exponential function  $f(t) = a \cdot \exp(b \cdot t)$ , where  $a$  and  $b$  are fit parameters to the recovery window after the Earth-point, which is defined as the 150 cadences after an Earth-point gap. We use the curvature of the fitted exponential, calculated by the average of its second derivative in the recovery region,  $\frac{d^2 f(t)}{dt^2}$ , to quantify the strength of the exponential character. As with the other goodness metric components, the goodness shall be expressed as a value in the interval [0,1), which we achieve by

<sup>10</sup>As the residuals are typically only very poorly correlated.

the regularization:

$$G'_{EP} = \frac{1}{1 + \alpha_{EP} \cdot \frac{d^2 f(t)}{dt^2}}, \quad (8.7)$$

where  $\alpha_{EP}$  is a weighting parameter to adjust the relative emphasis of this component in the total goodness metric  $G$ , which is the geometric mean of the four components. We also tried other metrics to quantify the Earth-point correction, including measuring the deviation of the 150 cadences window after an Earth-point from an auto-regressive estimate of this window, and another similar approach to the current one in which we also compare to an exponential fit of the region before the Earth-point gap and after the recovery window for reference. Out of the different metrics we tried, the current one is most in line with the perceived correction quality found upon manual investigation.

The above four components increase in value (in the range of  $[0, \infty)$ ) as the performance becomes more poor. But we wish for the “goodness” to vary between 0 and 1, 1 being perfect goodness, and so each component is inverted:

$$G'_{k,i} = \frac{1}{G_{k,i} + 1}. \quad (8.8)$$

The total goodness is then the geometric mean of the four components for residual correlations ( $G_C$ ), noise ( $G_N$ ), stellar variability ( $G_V$ ), and Earth-point recoveries ( $G_{EP}$ ):

$$G = \sqrt[4]{G'_C \cdot G'_N \cdot G'_V \cdot G'_{EP}}. \quad (8.9)$$

The four weighting parameters are used to adjust the relative emphasis of the four components. They have been empirically tuned to agree with observations of cotrending performance, with the weighting parameters being  $\alpha_C = 12.0$ ,  $\alpha_N = 1.0 \times 10^{-4}$ ,  $\alpha_V = 1.0 \times 10^4$ , and  $\alpha_{EP} = 5.0 \times 10^6$ .

The above computed goodness metric has been experimentally developed. It is used for target vetting, when reviewing data products, and as an aid to data users to give an estimate of how well to trust the cotrending performance.

### 8.3.4 Outputs of PDC

The output of PDC is a data structure containing the corrected light curves. If Harmonics Removal was performed (Subsubsection IV), a second set of corrected light curves without harmonic content is exported as well. In addition to these flux time series, PDC outputs data processing details such as the target and cadence indices of all SPSDs that were identified, the location of outliers, and diagnostic information such as the Bayesian prior weighting for each target. Of these data, TPS uses the corrected flux time series with filled gaps. The corrected flux times series, cotrending basis vectors, and additional processing data such as the goodness metric for each target are exported to the MAST.

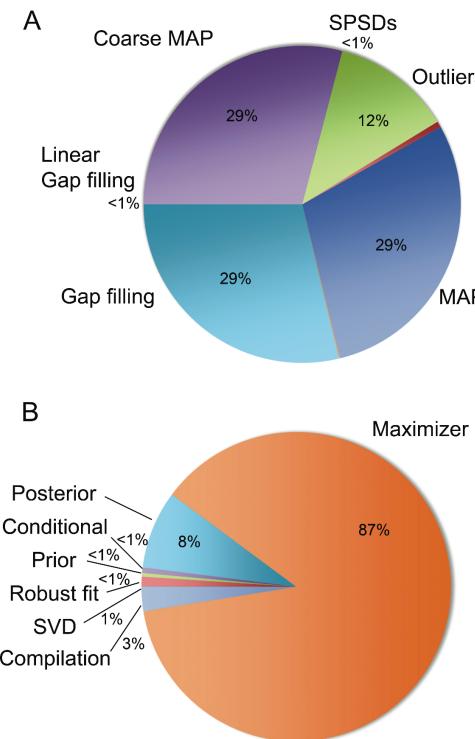
### 8.3.5 Processing Times

The unit of work for PDC (long cadence) is one channel per quarter, typically 1000–3500 targets with ~4500 data points each. Processing of *Kepler* data is performed on the *Kepler* SOC computer clusters and on the NASA Advanced Supercomputing Division’s Pleiades supercomputer at NASA Ames Research Center. The relative processing times of the different PDC components is displayed in Figure 8.15A and shows that the MAP-fit and the final gap filling are the computationally most expensive operations. Within MAP (Figure 8.15B), the maximization of the posterior PDF consumes the most time. Processing of a typical module output with 1,000–3,500 targets with PDC-MAP takes between 20 and 30 hours on a modern 3 GHz CPU with four cores.

NOTE: The processing times in this section are from SOC 8.0, when the proper gap filling was still performed in PDC and there was no multi-scale MAP. Processing times have since changed since this analysis. These figures are kept since they provide some information on the relative computational time of PDC components.

**8.3.5.1 A Note on Gap Filling** TPS requires contiguous data points for all cadences in a light curve, and moreover assumes statistical continuity of the wavelet coefficient variances of the flux time series. The simple linear gap filling operation (Subsubsection II) does not meet this requirement. Therefore, a more sophisticated gap filling algorithm is employed in TPS to refill the gaps of the corrected light curves. The filling occurs in TPS because it is very time consuming and TPS must have multi-quarter continuous time series. Filling the gaps properly in PDC would require a repeat of this very time consuming step in TPS anyway, so PDC produces light curves with the simple gap filling in Subsubsection II.

In TPS, the gap filler automatically distinguishes between gaps with “short” and “long” duration. The boundary for short and long gaps is determined by the gap filling module parameter set, and typically about 125 cadences. Gaps shorter than this threshold are filled with an autoregressive algorithm that estimates sample values in the gaps with a linear prediction based on the correlation



**Figure 8.15** Fractions of total processing time for PDC (panel A), and for the MAP part of PDC (panel B). From Figure 12 of Stumpe et al. (2012)

in the neighborhood of the gap. For gaps longer than this threshold, samples away from the edges would tend to become white noise samples, and a different algorithm is therefore used to estimate the gap values. Segments on either side of the gap are reflected onto the gap, and a weighted sum of the two segments is taken, where the weighting factor for each segment is a sigmoid function with the distance to respective edge of the gap. This procedure preserves statistical continuity of the data, and in particular, the correlation structure. However, some smoothing at small scales occurs due to the averaging of two noisy signals. To compensate for this effect, the variances of the wavelet coefficients of the filled gap are adjusted to match those of the original data in the neighborhood of the gap. Preserving the wavelet coefficient statistics is of particular relevance because TPS uses a wavelet-based approach for transit detection (Jenkins et al., 2010c, 2002; Tenenbaum et al., 2010).

## 8.4 A Bayesian Approach to Correcting Systematic Errors

The systematic errors observed in *Kepler* light curves exhibit a range of different timescales, from a few hours, to several days to many days and weeks. Such errors include, for example, temperature variations of the reaction wheel housing over the three-day momentum managements cycles and the resultant focus changes of  $\sim 2.2 \mu\text{m}$  per  $^{\circ}\text{C}$ . Large thermal effects can be observed in the science data for  $\sim 5$  days after recovering from intermittent safe modes, and for  $\sim 3$  days after attitude changes required to downlink the data each month, due to different sides of the spacecraft being heated during downlinks and subsequent thermal recoveries. Another prominent systematic error is DVA, which results in gradual trends in the data over each quarter. The principle objective of PDC is to remove these systematic effects by *cotrending*.<sup>11</sup> The fact that most systematic errors such as these affect all the science data simultaneously, though to differing degrees, provides significant leverage in dealing with these effects.

### 8.4.1 The Basic Problem and the Principle Behind the Solution

It is standard practice when removing systematic errors in stellar data to use robust LS on a set of basis vectors, exemplified in methods such as SYSREM (Tamuz et al., 2005) and TFA (Kovács et al., 2005). A robust LS approach, as outlined below in Subsection 8.4.2, can find a chance linear combination of the systematic error model components that reduces the bulk rms at the expense of distorting the intrinsic stellar variations and introducing additional noise on short timescales. The fundamental problem with this approach is the fact that the implicit model fitted to the data for each star is incomplete. Least-squares cotrending projects the data vector onto the selected basis vectors and removes the components that

<sup>11</sup>Detrending is the removal of arbitrary low-frequency signal content regardless of origin (intrinsic or systematic). In contrast, *cotrending* is the removal of specific signal content correlated across multiple targets, which can better preserve intrinsic low-frequency signals while removing wide-band systematic signals.

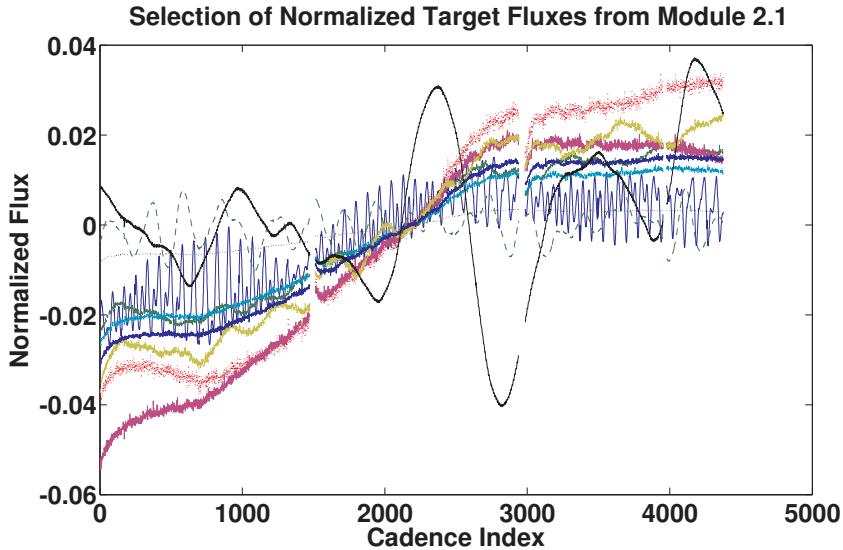
are parallel to any linear combination of the basis vectors. This process is guaranteed to reduce the bulk rms residuals, but may do so at the cost of injecting additional noise or distortion into the flux time series. Indeed, this occurs frequently for stars with high intrinsic variability, such as RR Lyrae stars, eclipsing binaries, and classical pulsators. For example, if one of the model terms is strongly related to focus variations and the long-term trend for the width of the stellar PSF is to broaden over the observation interval, then the flux for all stars should decrease over time. A LS fit, however, may invert the focus-related model term for a star whose flux increases over the observation interval, thereby removing the signature of intrinsic stellar variability from this light curve because there is a *coincidental* correlation between the observed change in flux and the observed change in focus. Given that the star would be expected to dim slightly over time, if anything, due to the focus change, PDC should be correcting the star so that it brightens slightly more than the original flux time series would indicate.

The situation is analogous to opening a jigsaw puzzle box and finding only 30% of the pieces present. LS gamely tries to put the jigsaw pieces together in order to match the picture on the box cover by stretching, rotating, and translating the pieces that were present in the box. The result is a set of pieces that roughly overlap the picture on the box cover, but one where the details don't necessarily match up well, even though individual pieces may obviously fit. In order to improve the performance of robust LS, we need to provide the fitter with constraints on the magnitudes and the signs of the fit coefficients. These constraints can be obtained by using the ensemble behavior of the stars to develop an *empirical* model of the underlying physics. For example, the photometric change that can be induced by a pointing change of 0.1 arcsec must be bounded, and this bound can be estimated by looking at how the collection of stars behaves for a pointing change of this magnitude.

As an example of this analysis and to demonstrate systematic trends in the *Kepler* data, take channel 2.1, the most thermally sensitive CCD channel in *Kepler*'s focal plane. Nearly all stars on this channel exhibit obvious focus- and pointing-related instrumental signatures in their pixel time series and flux time series. Figure 8.16 shows several characteristic light curves for typical targets<sup>12</sup> on channel 2.1 during Q7, normalized by the median flux value. Note the long-term increase for all flux curves over the 90-day interval. This is due to seasonal changes in the shape of the telescope and therefore its focus as the Sun rotates about the barrel of the telescope while the spacecraft orbits the Sun and maintains its attitude fixed on the FOV. All light curves exhibit these long term trends but to differing degrees. Also present are some short-term oscillations evident but mainly obscured in variable targets, due to focus changes driven by a heater cycling on and off to condition the temperature of the box containing reaction wheels 3 and 4 on the spacecraft bus. This component was receiving

<sup>12</sup>The light curves are referred to as “targets” and not stars since not all objects in the *Kepler* FOV are stellar. Galactic studies are also performed with *Kepler* data.

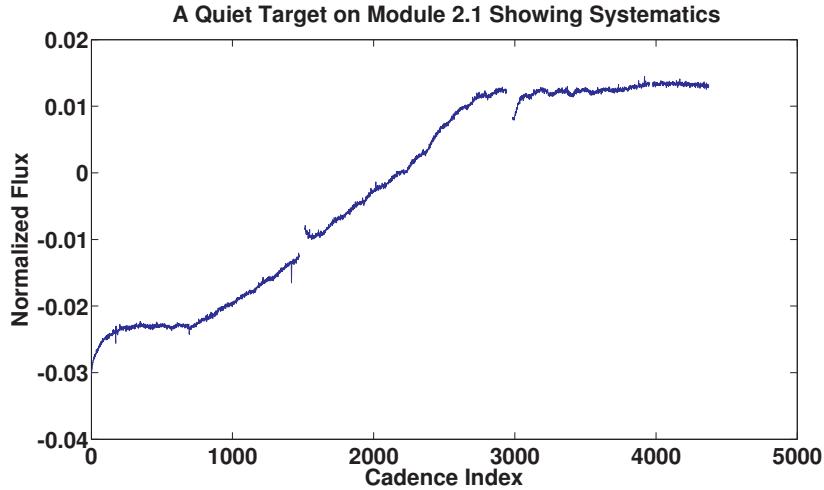
more and more shade throughout this time interval and the thermostat actuated more frequently over time. A target that is varying on levels and periods similar to these systematic effects can obscure the systematics making identification difficult. Looking at a single quiet target shown in Figure 8.17 (the same target



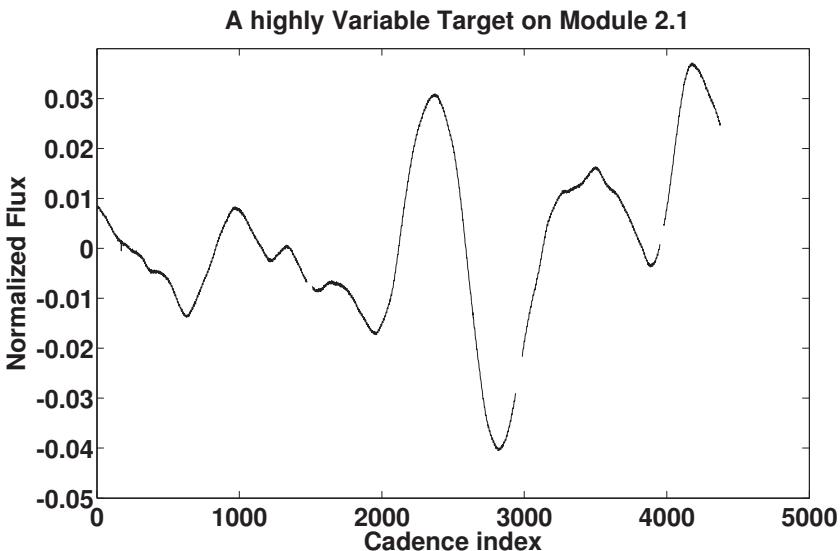
**Figure 8.16** Selection of typical light curves on channel 2.1. Notice the long-term trend exhibited in all light curves. However, for highly variable targets the trend is not entirely clear. This illustrates the need to separate intrinsic stellar variability from systematic trends. From Figure 1 of Smith et al. (2012).

shown in solid blue in Figure 8.16) allows us to more clearly see the systematic trends exhibited in all the targets in Figure 8.16 but obscured by variability. Each Earth-point, one at the beginning of the quarter (cadence index 0) and after each monthly downlink (cadences 1,500 and 2,800) results in a heating of different sides of the telescope as the spacecrafts re-orients the antennae to downlink data. The Earth-points themselves are gaps in the data that result in periods of local heating and cooling distorting the telescope. A characteristic recovery time is evident, as are other trends as described above. A final short data gap is also evident at cadence 3950, but this was not due to a reorientation of the spacecraft so no thermal recovery is present. As a counterexample, Figure 8.18 shows the same highly variable target in solid black in Figure 8.16. Notice how this highly variable star almost completely obscures the long-term trend. The targets shown in Figure 8.17 and Figure 8.18 will hereby be referred to as the *quiet target* and the *variable target* and used as canonical example targets in Section 8.5.

How can we separate intrinsic stellar variability from instrumental signatures? We do not expect intrinsic stellar variability to be correlated target to target, except for rare coincidences, and even then one would not expect a high degree of correlation for all timescales. However, we *do* expect instrumental signatures to be highly correlated from target to target and can exploit this observation to provide constraints on the cotrending that PDC performs.



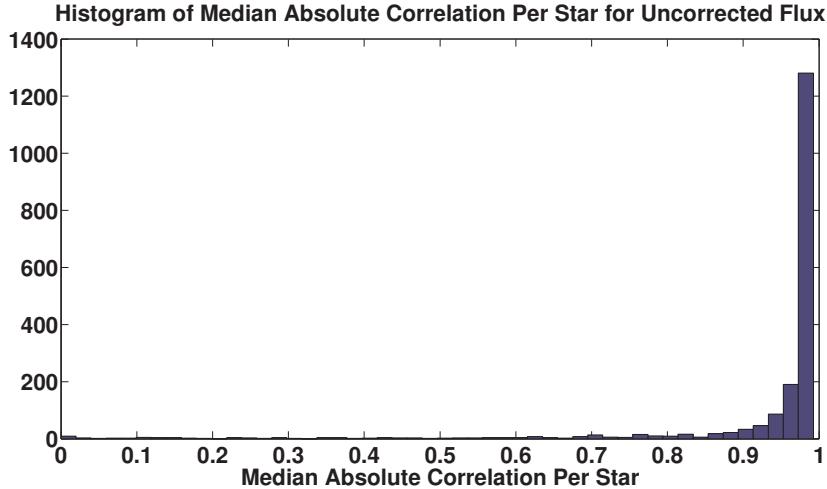
**Figure 8.17** A particularly quiet target on channel 2.1 showing almost purely systematic trends. The long-term trend is due to the seasonal changes to the shape of the telescope as the sun rotates around the barrel. Other spacecraft systematics are also visible such as monthly Earth-point downlinks and heater cycling. Data gaps and their thermal recoveries during the monthly downlinks are evident at cadences 1500 and 2800. From Figure 2 of Smith et al. (2012).



**Figure 8.18** A highly variable target where the variability completely obscures the systematic trend.

Figure 8.19 shows a histogram of the absolute value of the correlation coefficient for 1864 targets on channel 2.1. The targets' light curves are highly correlated as evidenced by the near complete pile-up near an absolute correlation coefficient of 1. Examination of individual light curves indicates that these light curves are contaminated to a large degree by instrumental signatures, as evidenced in Figure 8.16 and Figure 8.17. But not all the targets are dominated by systematic errors. The trick is to come up with a method that can distinguish

between intrinsic stellar variability and chance correlations with linear combinations of the diagnostic time series used to cotrend out systematic errors.



**Figure 8.19** Median Absolute Correlation for all targets on channel 2.1 for Q7.

#### 8.4.2 The MAP Approach, An Analytical Solution

The PDC-MAP method allows us to provide PDC with constraints on the fitted coefficients to help prevent overfitting and distortion of intrinsic stellar variability. In this exposition we follow the notation of Kay (2012).

The PDC-MAP technique examines the behavior of the robust LS fit coefficients across an ensemble of targets on each CCD readout channel in order to develop a description for the “typical” value for each model term. This description is a PDF that can be used to constrain the coefficients fitted in a second pass. To develop this approach, we build on a maximum likelihood approach.

The maximum likelihood approach models each light curve,  $\mathbf{y}$ , as a linear combination of instrumental systematic vectors, referred to as *Cotrending Basis Vectors* or CBV, arranged as the columns of a design matrix,  $\mathbf{H}$ , plus zero-mean, Gaussian observation noise,  $\mathbf{w}$ :

$$\hat{\mathbf{y}} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}. \quad (8.10)$$

The Maximum Likelihood Estimator (MLE) seeks to find the solution,  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ , that maximizes the likelihood function,  $p(\mathbf{y}; \boldsymbol{\theta})$ , given by

$$p(\mathbf{y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{C}_w|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}_w^{-1} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) \right], \quad (8.11)$$

where  $\mathbf{C}_w$  is the covariance of  $\mathbf{w}$  and  $N$  is the number of data points. Taking the gradient of the log of Equation 8.11, setting it equal to zero, and solving for  $\boldsymbol{\theta}$

yields the familiar LS solution:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = (\mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{y}. \quad (8.12)$$

This solution assumes the model  $\mathbf{H}$  is a complete model to the data. We will show that the Bayesian model accounts for an incomplete model, the common case when removing systematics from stellar signals.

Adopting the Bayesian approach allows us to incorporate side information, such as knowledge of prior constraints on the model, in a natural way. Bayesianists view the underlying model as being drawn from a distribution and the data as being one realization of this process. In this case we wish to find the MAP estimator of the model coefficients given the observations (data):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{y}) = \arg \max_{\boldsymbol{\theta}} p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad (8.13)$$

where we've applied Bayes' rule (D'Agostini, 2003) to simplify the expression. In this equation,  $p(\boldsymbol{\theta})$  is the *prior PDF* of the model coefficients. The mathematical form for  $p(\mathbf{y} | \boldsymbol{\theta})$  is the same as for the non-Bayesian likelihood function  $p(\mathbf{y}; \boldsymbol{\theta})$  in Equation 8.11.

For illustration purposes, if we adopt a Gaussian form for the coefficient distribution,  $\boldsymbol{\theta}$ , then  $p(\boldsymbol{\theta})$  takes a closed form solution:

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{M}{2}} |\mathbf{C}_{\boldsymbol{\theta}}|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \right], \quad (8.14)$$

where  $\mathbf{C}_{\boldsymbol{\theta}}$  and  $\boldsymbol{\mu}_{\boldsymbol{\theta}}$  are the covariance and mean of  $\boldsymbol{\theta}$ , respectively, and we assume that the coefficients are uncorrelated (which will hold true for orthogonal basis functions). We can then maximize Equation 8.13, using Equation 8.14, by maximizing its log likelihood:

$$\begin{aligned} \ln[p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})] &= \ln[p(\mathbf{y} | \boldsymbol{\theta})] + \ln[p(\boldsymbol{\theta})] \\ &= -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}_w| - \frac{1}{2} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}_w^{-1} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) \\ &\quad - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}_{\boldsymbol{\theta}}| - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}). \end{aligned} \quad (8.15)$$

Taking the gradient of Equation 8.15 with respect to  $\boldsymbol{\theta}$ , setting it to zero, and solving for  $\boldsymbol{\theta}$  yields

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = (\mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} (\mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{y} + \mathbf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}}). \quad (8.16)$$

If the observation noise,  $w$ , is zero-mean, white Gaussian noise with variance  $\sigma^2$ , then Equation 8.16 can be rewritten as

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = (\mathbf{H}^T \mathbf{H} + \sigma^2 \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} (\mathbf{H}^T \mathbf{y} + \sigma^2 \mathbf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}}). \quad (8.17)$$

The key to this Bayesian technique is to determine when to preference, or weight, the prior PDF over the conditional PDF. If the variance in the data is large compared to the “spread” allowed by the prior PDF for the model, then the MAP estimator gives more weight to the prior so that  $\hat{\theta}_{\text{MAP}} \rightarrow \mu_{\theta}$  as  $\sigma^2 \rightarrow \infty$ . This case would correspond, for example, to targets with large stellar variability such as those with the target given in Figure 8.18. In this case, the MAP weighting constrains the fitter from distorting the light curve and introducing noise on a short timescale. Conversely, if the variance in the data is small compared to the degree to which the prior PDF confines the model, the MAP estimator “trusts” the data over the prior knowledge and  $\hat{\theta}_{\text{MAP}} \rightarrow \hat{\theta}_{\text{MLE}}$  as  $\sigma^2 \rightarrow 0$ . This case would correspond to targets with small stellar variability such as with the target in Figure 8.17 where there is little risk of overfitting and distortion of the light curves and it is a “safe” bet to use the conditional, LS fit.

## 8.5 The Empirical Bayesian MAP Approach and Implementation

The above analytical solution to the Bayesian posterior PDF restricts the prior PDF to a Gaussian form. There is no *a priori* reason to make this assumption and in general, since we are developing an *empirical* prior PDF, the fewer analytical constraints on the form, the more complete will be the empirical model.

If a Gaussian form to the prior is no longer assumed then the prior formalism in Equation 8.14 can no longer be used. We can, however, still take the log form of Equation 8.13 to obtain

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta|y) = \arg \max_{\theta} (\log(p(y|\theta)) + \log(p(\theta))). \quad (8.18)$$

Using the MLE in Equation 8.11 for  $p(y|\theta)$ , removing the constant terms, inserting a weighting parameter, and using normalized light curves,  $\hat{y}$ , we obtain

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \left[ -\frac{1}{2\sigma^2} (\hat{y} - \mathbf{H}\theta)^T (\hat{y} - \mathbf{H}\theta) + \mathbf{W}_{\text{pr}} \log p(\theta) \right], \quad (8.19)$$

where we assume the observation noise,  $w$ , is zero-mean, white Gaussian noise and has variance  $\sigma^2$ . Since  $p(\theta)$  is no longer in closed form, the “spread” in the prior PDF (i.e., the covariance of  $\theta$ ,  $C_{\theta}$  in Equation 8.17) can no longer be expressed succinctly. In its stead, a *generalized weighting parameter*,  $\mathbf{W}_{\text{pr}}$ , is used to characterize the “spread” in the prior PDF. Equation 8.19 must now be evaluated numerically.

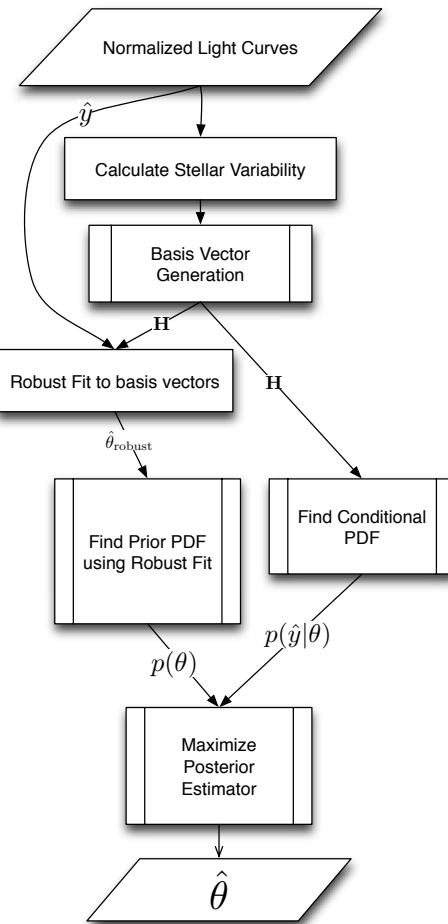
The overall flow of the algorithm is shown in Figure 8.20. We start by normalizing the flux light curves and calculating a relative stellar variability. We then find basis vectors using SVD based on a reduced set of flux light curves where cuts are made on target-to-target correlation and stellar variability. A robust LS fit is then performed on each target using these basis vectors. This ensemble of *robust fit coefficients* is used to generate the prior PDF. The conditional PDF is

also found based on the same basis vectors. Once both prior and conditional PDF are found they are combined to generate the posterior PDF, where a weighting parameter, based on the stellar variability and the “goodness” of the prior fit, is used to weigh the prior relative to the conditional PDF. Details are elucidated in the following subsections.

### 8.5.1 Finding the Cotrending Basis Vectors

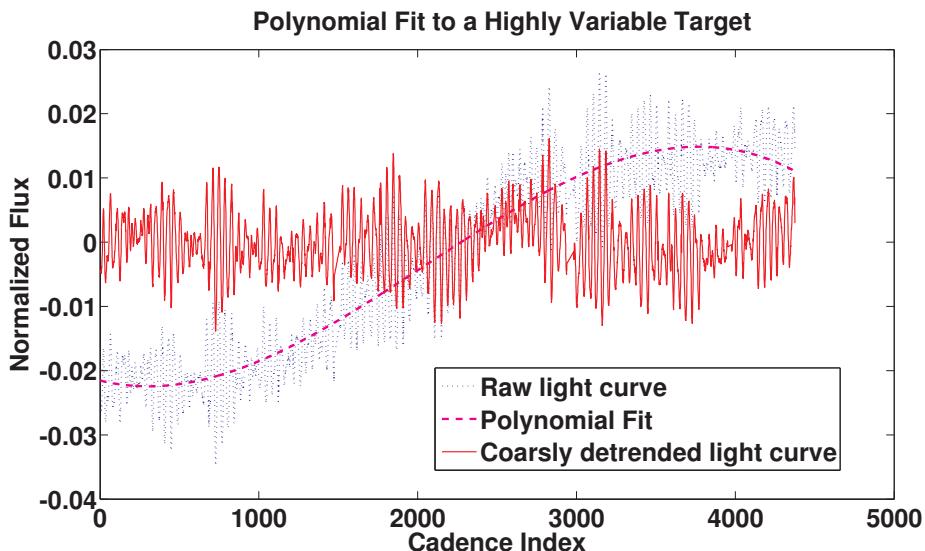
The CBVs are obtained using SVD. In order to have equal representation for all light curves independent of their absolute magnitude, we first normalize the targets by their mean flux values ( $\frac{\Delta\text{flux}}{\text{mean}(\text{flux})}$ ). We then select the 50% most highly correlated targets based on the median absolute Pearson correlation. This cut generates a set that exhibits the strongest trends in the data. It mostly removes targets with large variability but not completely. A variable star exhibiting a strong trend can still remain in the reduced list. We therefore first make a cut on the estimated variability of each target.

An estimate of the intrinsic stellar variability of each target must be found. Herein lies the fundamental chicken-and-egg problem of the cotrending method. We need to know the stellar variability of each target in order to know how much to rely on the prior. But if we already knew the stellar variability then we would have no need for the prior – the cotrending solution would simply be the intrinsic stellar variability subtracted from the light curve, minus a Gaussian noise estimate. This issue is not specific to this particular cotrending method either. Whenever a system is characterized with an incomplete model there exists the problem of identifying the components in the system not represented in the model. We fortunately do not need to absolutely know the variability; we only need an estimated metric in order to weigh the prior. This estimate can be obtained by comparing a third-order polynomial to the light curve. The polynomial



**Figure 8.20** Flowchart of the PDC-MAP cotrending algorithm.

will remove any long-term trends leaving behind a *roughly* detrended curve. The standard deviation of this polynomial removed light curve results in a rough calculation of the variability of the target. Removing a low-order polynomial is essentially a high-pass filter; we are therefore assuming any long-term trends are systematic and short-term trends are stellar. There are numerous counterexamples of short-term trends that are actually systematic – reaction wheel zero crossings is a good example. However, short-term systematic trends tend to be small in magnitude whereas long-term systematics tend to result in large diversions in the flux amplitude. Likewise, there are examples of intrinsic long-term trends but they are generally smaller than the systematic trends. Since we are only concerned with the *relative* amplitude of stellar versus systematic variability, we are using the low-pass filter to distinguish two characteristic realms of influence: long-term trends dominated by systematics and short term trends dominated by intrinsic stellar variation. An example is shown in Figure 8.21. Here, a highly variable target is compounded with a long-term DVA and thermal trend. For periods less than 400 cadences the variance in the flux is dominated by stellar features. The long-term variance and the general trend to higher flux values are due to systematics. The variance of the residual after removing the polynomial fit, labeled as “Coarsely detrended light curve” in the figure, gives a *rough* estimate of the stellar variability of this target. Note that there are still systematic features in the detrended light curve. They are, however, small in magnitude compared to the stellar variability<sup>13</sup>.



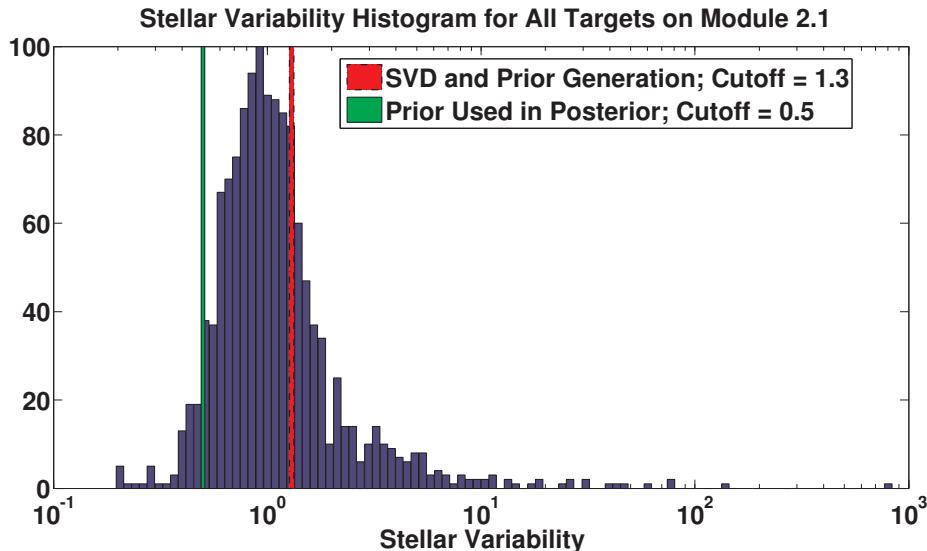
**Figure 8.21** By removing a third-order polynomial fit to the raw light curve an estimate of the intrinsic variability of the target can be calculated. From Figure 6 of Smith et al. (2012).

<sup>13</sup>This does not lessen the ability of PDC-MAP to remove short-term systematics. Such short-term systematics are still present in the basis vectors and so when the PDF fit is performed the short-term trends are removed.

The variability,  $V$ , is measured using

$$V = \frac{\sigma_{\hat{y}}}{\Delta y \tilde{V}}, \quad (8.20)$$

where  $\sigma_{\hat{y}}$  is the standard deviation of the third-order polynomial detrended light curve,  $\Delta y$  is the uncertainty of the flux data as determined by the PA pipeline component (Clarke et al., 2010) and  $\tilde{V}$  is the median variability over all light curves in the sample. The normalization by the uncertainty is to ensure the noise in the data is not included in the stellar variability. The normalization by the median variability is so that a variability of 1 is considered typical, thereby simplifying the analysis parameterization. Figure 8.22 shows a histogram of the measured variability for all targets on channel 2.1. The median of all of these values is evidently 1 and the distribution is typical for all channels, where most are close to typical variability but with a long tail to high variability (note the log scale for the x-axis). There are two cutoff thresholds plotted as well. The upper (in dashed red) is the threshold to determine if a target is “highly variable.” The lower (in solid green) is to determine if a target is “very quiet.” The very quiet targets have such a low amount of variability that using the prior PDF when generating the fit has been found to be problematic. Any targets above the high



**Figure 8.22** Histogram of estimated variability for all targets on channel 2.1. This distribution is typical for all *Kepler* channels. The quiet targets below the “SVD and Prior Generation” threshold are used to generate the cotrending basis vectors. From Figure 7 of Smith et al. (2012).

variability threshold are removed from the reduced list. The remaining targets are sorted with respect to median absolute correlation and the 50% most highly correlated are used for SVD.

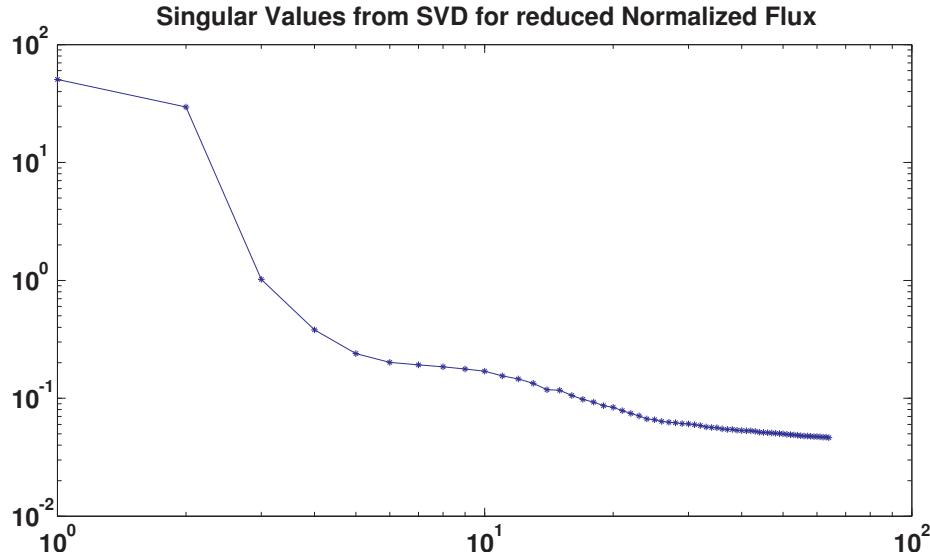
It should be noted that if the targets are normalized by their median before SVD then most targets pass through zero amplitude at the midpoint, as can be

seen as a “node” in the light curves at cadence 2200 in Figure 8.16. If SVD was performed on this set, as is, then all the strong cotrending basis vectors would have zero amplitude at the midpoint. The basis vectors would therefore be unable to remove systematics in the minority of targets that do not pass through zero at the midpoint. The light curves could be *dithered* slightly by a zero-mean Gaussian dithering magnitude in order to slightly “spread” the light curves about the zero flux value. Since the dithering is zero-mean, this has no effect on the resultant basis vectors other than to remove the artificial zero-crossing node at the midpoint. Note that the dithering would only be used to generate the basis vectors. The cotrending would still be performed on the non-dithered, but median-normalized, light curves. An alternative method, and the method currently used in PDC is to *mean* normalize the light curves for SVD and basis vector generation. This removes the “node” midway through the quarter and the need for dither. However, *median* normalized light curves are still ideal for the cotrending due the better regularization of the light curves. So, the cotrending is still performed on median normalized light curves. The absolute magnitude of the basis vectors is irrelevant during cotrending and we find no conflict in having the light curves normalized in two different ways for each step.

Figure 8.23 shows the singular values from the singular value decomposition. This figure is characteristic of all channels: two or three strong singular values, then a slowly tapering region for about another dozen values until finally asymptotically approaching zero (as is expected with SVD). The first several left singular vectors are selected (typically the first eight) to become the *Cotrending Basis Vectors*. These first singular vectors exhibit the principle trends in the data due to DVA, pointing errors, impulses due to Argabrightenings (Witteborn et al., 2011), focus errors and reaction wheel zero crossings among other trends. The number of basis vectors used is generally eight; however a signal-to-noise ratio (SNR) test is performed, where the SNR is determined by

$$\text{SNR}_{\text{db}} = 10 \log_{10} \left( \frac{A_{\text{signal}}^2}{A_{\text{noise}}^2} \right). \quad (8.21)$$

$A_{\text{signal}}$  and  $A_{\text{noise}}$  being the rms of the light curve and noise floor, respectively. The noise floor is approximated by the first differences between adjacent flux values. Any of the eight basis vectors with a SNR below a threshold of 5 decibels are removed but only a small number of basis vectors over the entire field of view are removed by the SNR test. Most have high SNR. A second dimensionality method is used based on Bayesian Model Selection (BMS, Minka, 2008). In most cases this method will pick too high a dimensionality so we always pick the lesser number of basis vectors given by either the SNR or BMS test. We wish to find only the singular vectors with systematics; the lesser singular vectors do contain light curve signal information, but not necessarily systematics. The Bayesian prior performs well at reducing the application of these lesser basis vectors but we have found including them in the MAP fit adds no value yet slows down the algorithm.



**Figure 8.23** The singular values from SVD on the reduced set of “cleaned” targets that are highly correlated and quiet. From Figure 8 of Smith et al. (2012).

For a minority of basis vectors, a few target light curves can dominate the signal. The normalization process attempts to “equalize” the strength of all targets, but a small number of light curves can be over-represented in the singular vectors from SVD. To eliminate this we calculate an entropy metric for each basis vector using the following entropy calculation:

$$h(p_i) = - \int p(x) \log p(x) dx, \quad (8.22)$$

where  $p(x)$  is a probability distribution function created from the right singular vectors from SVD (referred to as the *V-Matrix*),

$$p_i(x) = \{V_{ki}\}. \quad (8.23)$$

The V-Matrix contains the contribution of the signal in the basis vector from each target light curve. We must first normalize the entropy calculation to a Gaussian distribution, which has the highest entropy of any continuous distribution with the same 2nd moment. The entropy of a Gaussian is

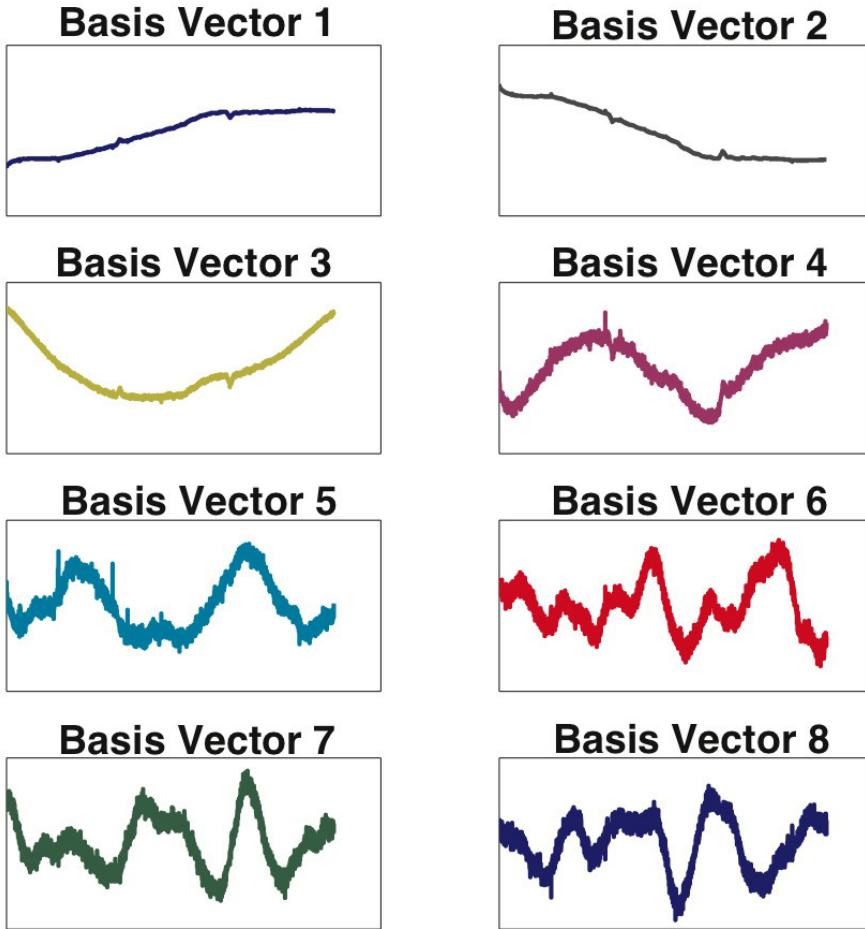
$$H_0(\sigma) = \frac{1 + \log(2\pi)}{2} + \log(\sigma), \quad (8.24)$$

$\sigma$  being the 2nd central moment of  $V_{ki}$  for fixed  $i$ . The resultant relative entropy is therefore

$$h'(p_i) = h(p_i) - H_0(\sigma). \quad (8.25)$$

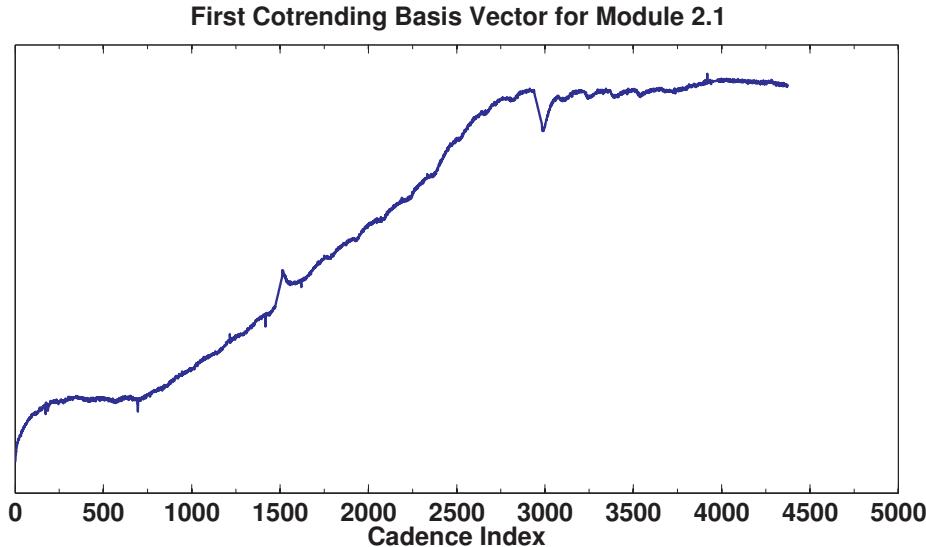
If one (or a few) targets dominate, then they will have much larger values in the V-Matrix than all the other targets. A negative value of the entropy calculation

will identify this condition. Bad entropy is somewhat arbitrary, but we have found that a value below  $-0.7$  is poor. For any basis vectors with identified poor entropy, the V-matrix column for that basis vector is examined for standout targets. The offending targets are removed and SVD is re-computed on the remaining targets. The process is iterated until the entropy of all basis vectors is below  $-0.7$ . Typically, no more than a couple of iterations is necessary and fewer than 20 targets are removed (out of 2,500 total targets).



**Figure 8.24** First eight Cotrending Basis Vectors for channel 2.1. The gaps in the data have been linearly filled so these curves are continuous. From Figure 9 of Smith et al. (2012).

Figure 8.24 shows the first eight cotrending basis vectors generated for channel 2.1 and Figure 8.25 shows just the first basis vector. Trends can be found in all the vectors but it is useful to concentrate on the first and strongest. Here the most characteristic trends and systematics in the data can be found. The general trend toward higher flux is due to seasonal change and solar orientation. The short recovery periods at cadence indices 0, 1500, and 2800 are due to monthly



**Figure 8.25** First Cotrending Basis Vector for channel 2.1. The amplitude of the basis vector is arbitrary. From Figure 10 of Smith et al. (2012).

downlinks. The short spikes at 700 and 1450 are due to artifacts from correcting cosmic rays near reaction wheel zero-crossing periods.<sup>14</sup> The periodic oscillation is due to heater cycling. Notice how the basis vector in Figure 8.25 closely follows the flux light curve in Figure 8.17. This signifies that virtually all the features in this flux light curve are due to systematic effects and not intrinsic stellar variability. In theory, any features in the light curves in Figure 8.16 that are not represented in the basis vectors in Figure 8.24 are intrinsic to the target. However, a simple LS projection of the light curves on the basis vectors will not produce desirable results for all targets as will be shown below.

Once the cotrending basis vectors are found a robust LS fit is performed on each target. This creates the empirical data used to generate the prior PDF.

### 8.5.2 Numerically Generating $p(\theta)$

The prior PDF is based on the distribution of robust fit coefficients of the basis vectors for all light curves using the method described in Bowman & Azzalini (1987). This method computes a probability density estimate of the sample data based on a normal kernel function using a window that is a function of the number of points in the data sample. The form of the prior PDF will depend on the parameterization of the robust fit coefficients. We must thus decide how to parameterize the coefficients to best extract the correlations. Some systematic effects are caused by focal plane irregularities, others by thermally induced focus changes, which are stronger near the edges of the CCD frame due to the

<sup>14</sup>These artifacts have been resolved in a recent version of the PA pipeline component but Argabrightenings still persist.

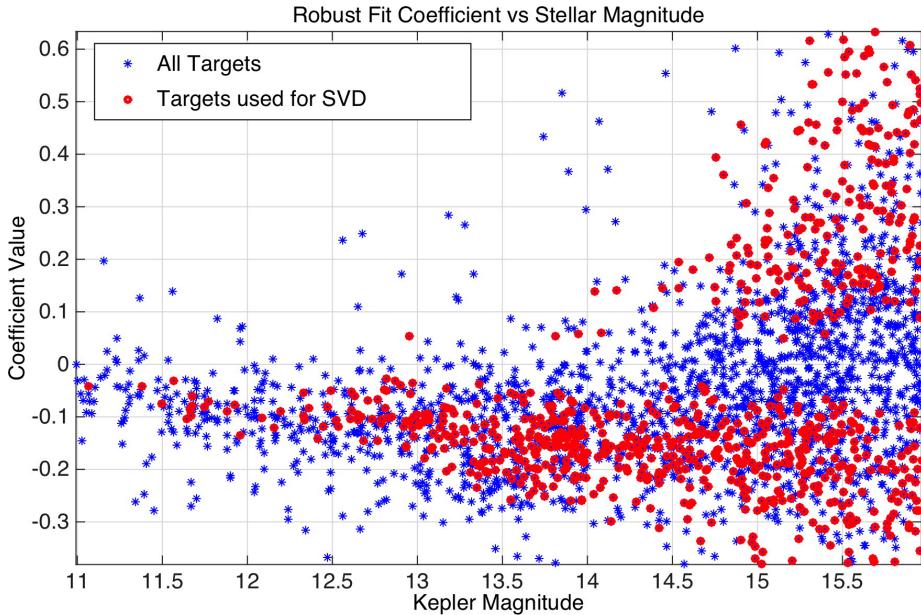
outer edges being further away from the optical ring of best focus. There are also other issues that are dependent on the physical position of each pixel on the CCD. Therefore, the targets' locations in the sky as characterized by RA and Dec are reasonable parameters to characterize target location with respect to the sources of systematic effects. The targets' influence by systematic effects is also directly related to the stellar magnitude since different magnitude targets result in different saturation levels of the CCD pixels. For example, the readout electronics for the CCD are sensitive to temperature drift but the sensitivity is non-linear with respect to CCD flux levels. So brighter targets are affected by instrument temperature differently than dim targets. We therefore parameterize the prior PDF with three independent variables: 1) stellar magnitude ( $K_p$ ), 2) right ascension (RA), and 3) declination (Dec).

Figure 8.26, Figure 8.27, and Figure 8.28 show the robust fit coefficients for a basis vector plotted against  $K_p$ , RA, and Dec. The blue star data is for all targets whereas the red circle data are just for those targets remaining for SVD after the cuts discussed in Subsection 8.5.1. The solid blue and dashed red curves in Figure 8.27 and Figure 8.28 are the traveling window means of the blue star and red circle data, respectively. The cuts clearly produce a bimodal distribution in  $K_p$  for this basis vector. A simple Gaussian fit would not reproduce this, demonstrating that the systematic trends are correlated with  $K_p$  but variable targets are masking the true correlation when a simple robust fit is performed. The correlations in RA and Dec are also evident but to a lesser extent. Notice also that the mean (solid blue curves) are biased compared to the dashed red. This is again because the variable targets are masking the true trends in the data. Some basis vectors exhibit stronger trends in  $K_p$ , RA, or Dec but not necessarily all three simultaneously, as is expected if the different systematics represented by the basis vectors have different instrumental sources. Plotting different basis vectors and/or channels reveals different trends and correlations.

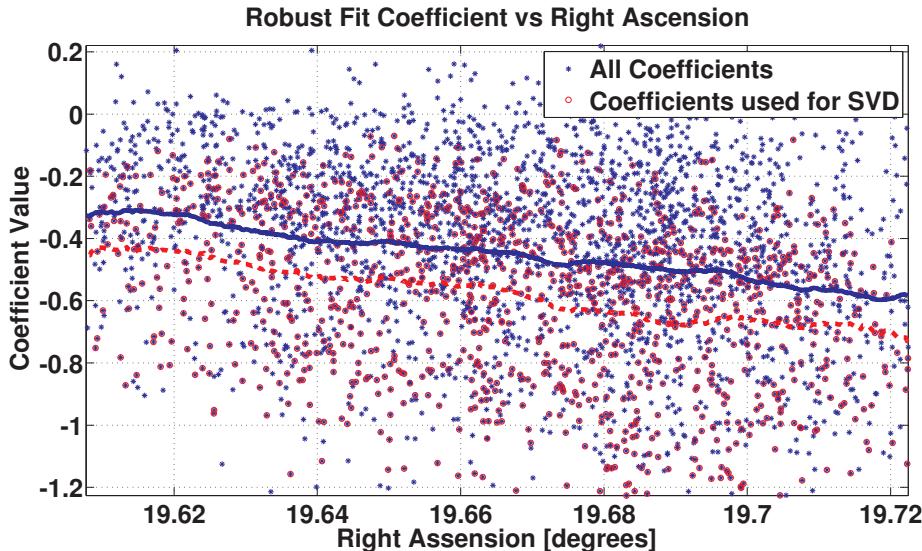
We want to mainly rely on targets in the neighborhood around the target that we are fitting, referred to as the *target under study* (TUS), in RA, Dec, and  $K_p$  space when generating the prior PDF. If we simply found an evenly weighted PDF then a large cluster of targets with a certain coefficient value, even if non-local to the TUS, would always dominate the peak of the prior PDF. We therefore use a weighted probability density estimate based on the standard Euclidean distance between targets  $\mathbf{x}$  and  $\mathbf{y}$ :

$$D = \sqrt{(\mathbf{x} - \mathbf{y}) \Lambda^{-1} (\mathbf{x} - \mathbf{y})^T}, \quad (8.26)$$

where  $\Lambda$  is a diagonal matrix whose diagonal elements give the relative weighting for each dimension. A straight normalization in each dimension by its standard deviation would result in equal weighting of all three dimensions, but we wish to accentuate the prior PDF in dimensions that exhibit greater correlations, and in our case, the robust fit coefficients exhibit a stronger correlation in  $K_p$ .



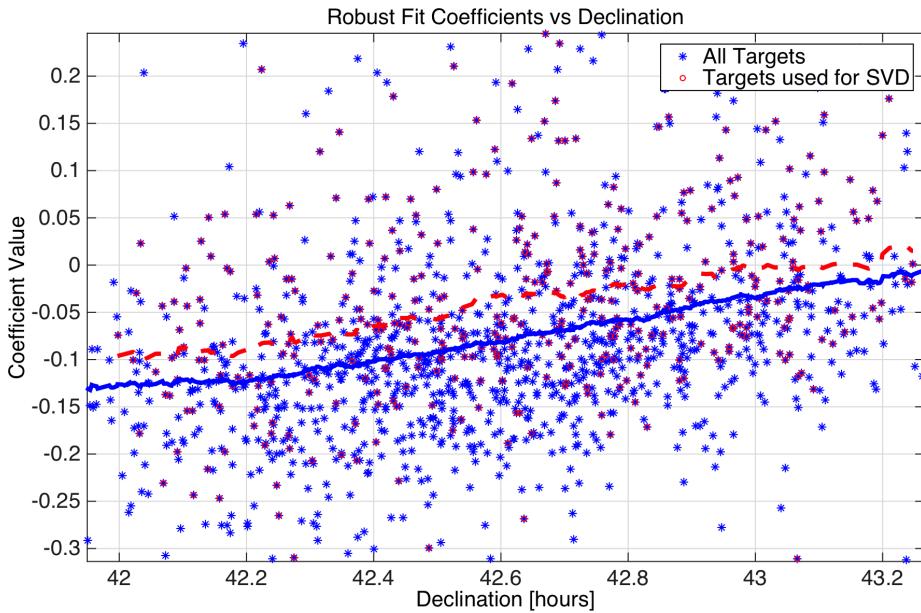
**Figure 8.26** Robust fit coefficients for Basis Vector 1 for all targets (blue stars) and only those targets used for SVD (red circles) plotted against *Kepler* magnitude. Taking cuts on stellar variability, target-to-target correlation, and entropy results in a bimodal distribution that would not be evident without the cuts. From Figure 11 of Smith et al. (2012).



**Figure 8.27** Robust fit coefficients for Basis Vector 1 for all targets (blue stars) and only those targets used for SVD (red circles) plotted against Right Ascension. From Figure 12 of Smith et al. (2012).

than in RA or Dec. The  $\Lambda$  matrix diagonals are therefore:

$$\Lambda_i = \frac{\text{mad}(\boldsymbol{\theta}_i)}{S_i}, \quad (8.27)$$

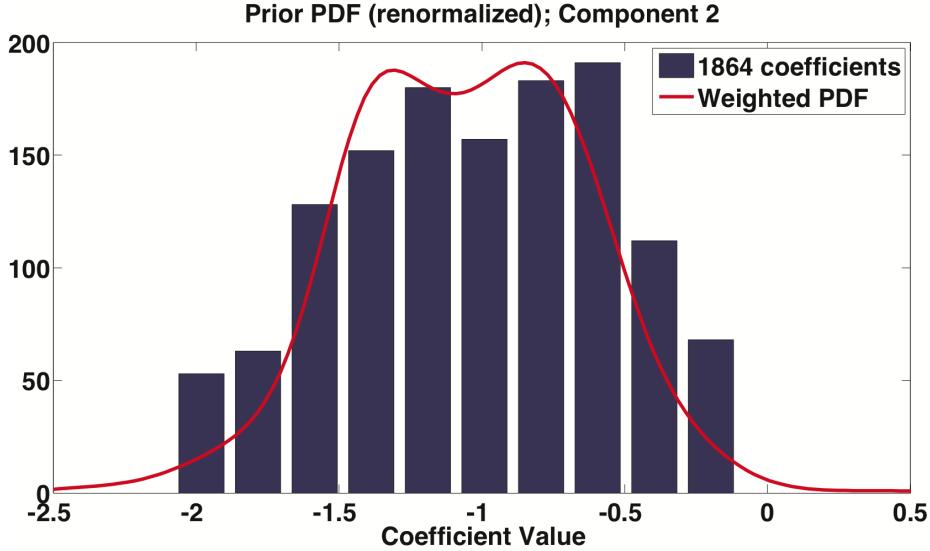


**Figure 8.28** Robust fit coefficients for Basis Vector 1 for all targets (blue stars) and only those targets used for SVD (red circles) plotted against Declination. From Figure 13 of Smith et al. (2012).

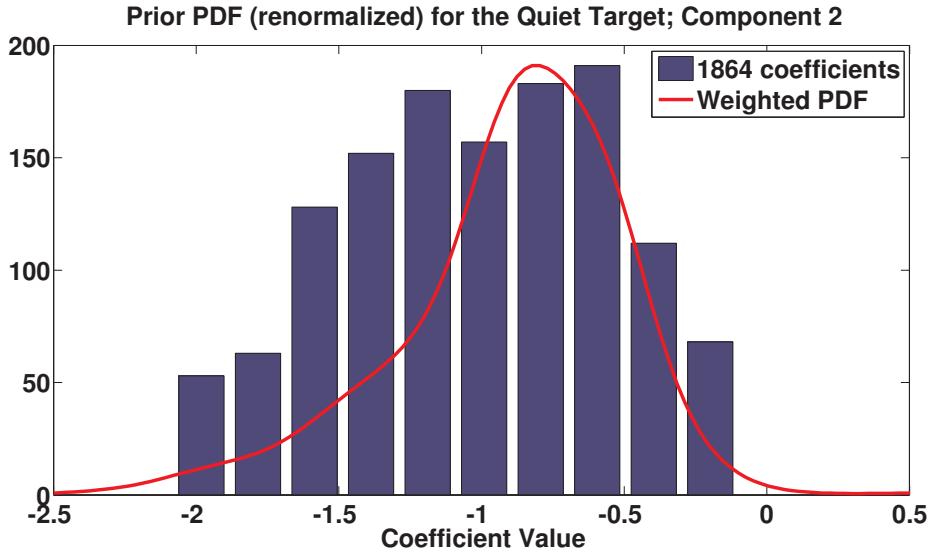
where  $\text{mad}(\theta_i)$  is the median absolute deviation of the coefficient distribution along dimension  $i$ , and  $S_i$  is the scaling factor for dimension  $i$ . ( $S_i = \{2, \text{if } i \Rightarrow K_p \text{ or } 1, \text{otherwise.}\}$ ) The above weighting results in the  $K_p$  dimension weighted twice as much as RA and Dec when generating the prior PDF. That is, targets farther away in the  $K_p$  dimension are weighted proportionately less than in RA and Dec. This effectively results in taking a tighter cut in  $K_p$  space to emphasize the greater correlation in that dimension. Since the PDF is weighted by this distance metric, the PDF will emphasize the correlation in  $K_p$  and yet still be sensitive to the trends in RA and Dec. The median absolute deviation is used instead of the standard deviation in order to ignore outliers.

The weighting by Equation 8.26 and how it affects the prior PDF is illustrated in Figure 8.29, Figure 8.30 and Figure 8.31, the latter two being the prior PDFs for the same two targets in Figure 8.17 and Figure 8.18. The blue histogram in all three figures is exactly the same since they are generated from the same distribution of coefficients. However, the prior PDF (red curve) is dramatically different. In Figure 8.29 a bimodal PDF is evident due to targets nearby to the TUS containing two clusters around -1.3 and -0.85, which suggests that the coefficient value for the TUS should be one of these two values. The value that is actually chosen will be dependent on the form of the conditional PDF and the weighting of the prior PDF as discussed in Subsection 8.5.3. In Figure 8.31 the targets near the TUS have coefficients clustered around -0.34, which is far from the peak in the unweighted PDF. Using the unweighted PDF would have completely missed the actual systematic trend in the data near the TUS. The log of the prior PDF is

plotted in these figures for direct comparison with Equation 8.19, which results in a compression of the PDF near the top.

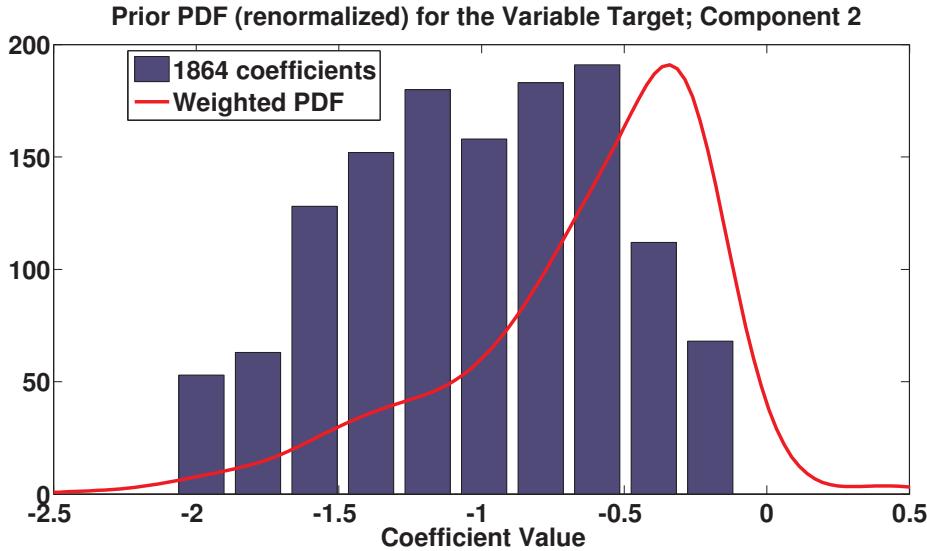


**Figure 8.29** Histogram of Basis Vector 2 robust fit coefficients for all 1,864 targets on channel 2.1 and the weighted probability density for a particular target. The weighting by distance in  $K_p$ , RA, and Dec clearly affects the PDF. From Figure 14 of Smith et al. (2012).



**Figure 8.30** Histogram of Basis Vector 2 robust fit coefficients for all 1,864 targets on channel 2.1 and the weighted probability density for the Quiet Target shown in Figure 8.17. From Figure 51 of Smith et al. (2012).

In summary, the prior PDF is developed by generating a 3-D weighted distribution of robust LS fit coefficients in RA, Dec, and  $K_p$  space. This methodology makes no assumptions, Gaussian or otherwise, on the form of the PDF, and al-



**Figure 8.31** Histogram of Basis Vector 2 robust fit coefficients for all 1,864 targets on channel 2.1 and the weighted probability density for the Variable Target light curve shown in Figure 8.18. From Figure 16 of Smith et al. (2012).

lows PDC-MAP to identify and characterize the form of the systematic trends across the full distribution of targets.

### 8.5.3 Finding the Weighting Parameter $W_{\text{pr}}$

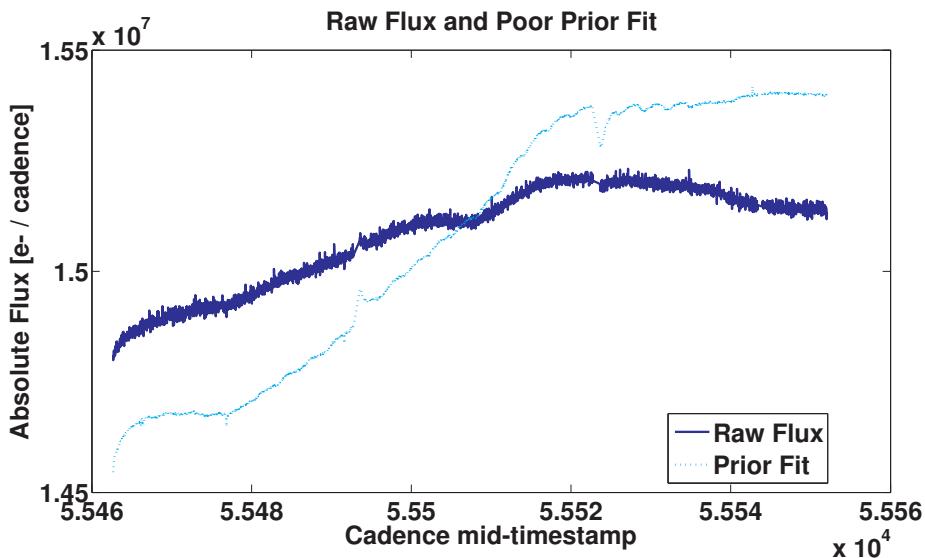
For each light curve the weighting parameter,  $W_{\text{pr}}$ , in Equation 8.19 is an empirical weighting parameter that is principally based on the variability of each target. The greater the variability, the greater we need to constrain the LS fit. However, there is another complication. One factor influencing the “goodness” of the prior fit is the sparseness of the targets in certain regions in RA, Dec, and  $K_p$  space. A sparse distribution will result in poor prior statistics. There are also other unknown causes resulting in poor priors for some targets, and so an additional parameter in the prior weighting is an evaluation of the “goodness” of the prior fit. The goodness is evaluated using a method similar to the variability calculation above. The prior fit is compared to a third-order polynomial fit to the light curve with a soft-wall cutoff using the following equation:

$$G_{\text{pr}} = \begin{cases} 1 - \left( \frac{G_{\text{raw}}}{\alpha_G} \right)^3, & \text{if } G_{\text{raw}} < \alpha_G \\ 0, & \text{otherwise} \end{cases} \quad (8.28)$$

where  $G_{\text{raw}}$  is the “raw” goodness given by

$$G_{\text{raw}} = \text{std} \left( \frac{(F_{\text{pr}} - F_{\text{poly}})}{\text{mad}(y - F_{\text{poly}})} - 1 \right), \quad (8.29)$$

and  $F_{\text{pr}}$  and  $F_{\text{poly}}$  are the prior PDF fit and the third-order polynomial fit to the data respectively. Normalization by the median absolute deviation (mad) of the polynomial fit removed light curve allows for a comparison of the difference between the polynomial fit and the prior fit with respect to the variance of the target. The soft cutoff is to ensure that small changes in the light curve will not have dramatic changes in the weighting. The scaling parameter  $\alpha_G$  is determined by when the deviation of the prior fit to the polynomial fit becomes too poor to be useful in constraining the *a posteriori* fit. An example of a poor prior fit is given in Figure 8.32. Notice how both the long-term trend and the Earth-point recoveries are much larger in the prior fit than in the actual data. Examples such as this are in the minority, but frequent enough to require the additional test for prior goodness. It could be supposed that this target is trending downward, canceling out the upward trend of the prior fit. This is unfortunately not the case. Examination of Q6 and Q8 reveals that this target is not experiencing a general trend in Q7. The prior fit is indeed poor and should not be used to any large degree. The unfortunate side effect of the prior goodness test is that PDC-MAP



**Figure 8.32** An example of a poor prior PDF fit to the trend in the target. From Figure 17 of Smith et al. (2012).

is less sensitive to very long-term trends in the data. A true long-term trend in the data that cancels out the systematic trend can confuse the prior goodness metric that would interpret the fit as a bad prior. The only way to surely know the actual long-term trend is to examine multi-quarter data. Future versions of the PDC module may indeed provide this functionality.

The resultant full form to the prior weighting is:

$$\mathbf{W}_{\text{pr}} = V^{\beta_V} G_{\text{pr}}^{\beta_G} \quad (8.30)$$

the parameters  $\beta_V$  and  $\beta_G$  being scaling factors for the variability and prior goodness, respectively.

In cases where the prior goodness is near zero, the fit reverts to a reduced robust fit where the number of basis vectors is limited to just the first several (the default is four). A MAP fit features a large number of basis vectors that can be used. The prior PDF restricts the fit from drifting drastically in function space, searching the large set of basis vectors for a combination that reduces the bulk rms at the expense of distorting stellar features. If the prior cannot be used then there is no such restriction and the posterior PDF becomes a LS fit, so a more limited number of basis vectors must be used in order to constrain the fit. The first several basis vectors have very strong trends in most of the data and have low noise components so they are generally safe to use even with an unrestricted LS fit. It is also generally true that a target has a bad prior because it is quiet and any small deviation in the prior from a true trend is very noticeable; therefore the prior is neither necessary or desirable to use.

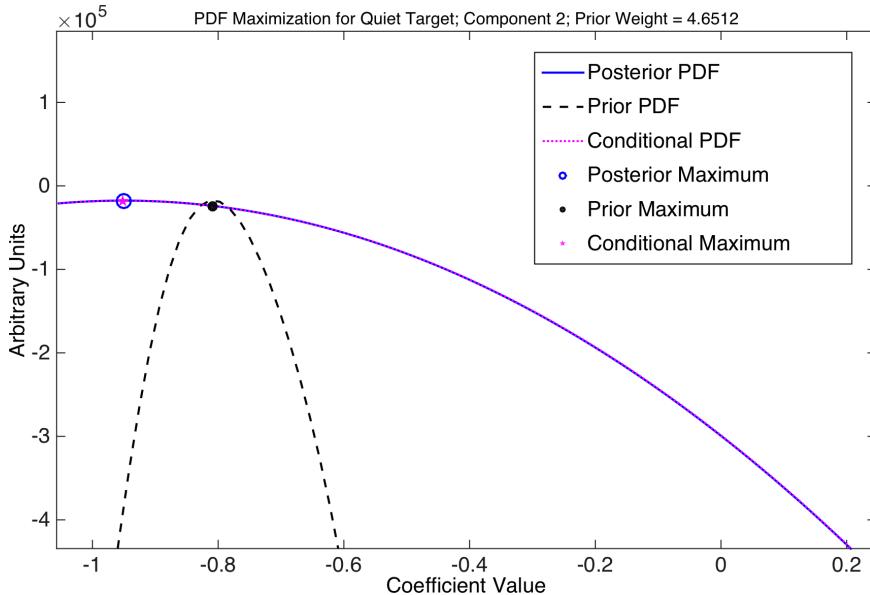
If the target is below the variability threshold shown in Figure 8.22 then the target is very quiet. In many cases the use of the prior fit only worsens the fit over a LS fit, so the prior weighting is zeroed. This is due to the prior fit never being an exact match to the target trend; even small deviations can “pull” the posterior fit away from a good fit. In such cases there is little risk of a quiet target biasing a LS fit away from a proper cotrending fit. The majority of targets do not fall into either of the above two cases, so the prior is used to the degree dictated by the prior weight and a Bayesian MAP fit is performed.

#### 8.5.4 Maximization of the Posterior PDF

Once the prior weighting is determined the posterior PDF can be assembled using Equation 8.19. Due to the empirical, and therefore non-analytical, form of the prior PDF, the posterior must be maximized numerically. In general, a multi-dimensional maximization is difficult and time consuming due to the risk of only finding local maxima. Fortunately, due to the use of SVD, the basis vectors are all orthogonal so the various coefficients  $\hat{\theta}_i$  can be maximized sequentially. The process is therefore straightforward. The strongest singular vector is maximized first and then all subsequent singular vectors are maximized in turn.

Following along with the same two examples of a quiet and a variable target, Figure 8.33 and Figure 8.34 show the final posterior PDF along with the prior and conditional PDF. The black dots and magenta stars are the maxima of the prior and conditional PDF and the blue circle is the maximum of the posterior PDF. Due to the varying scales of the three curves, for illustration purposes, the prior and conditional curves have been renormalized to the same scale as the posterior. The conditional fit has the smooth quadratic form characteristic of a LS fit. The prior appears Gaussian on this scale but in general is not. The title of each plot gives the prior weight. For the quiet target the variability is low so the prior weight is only 4.65, resulting in a minor correction to the conditional

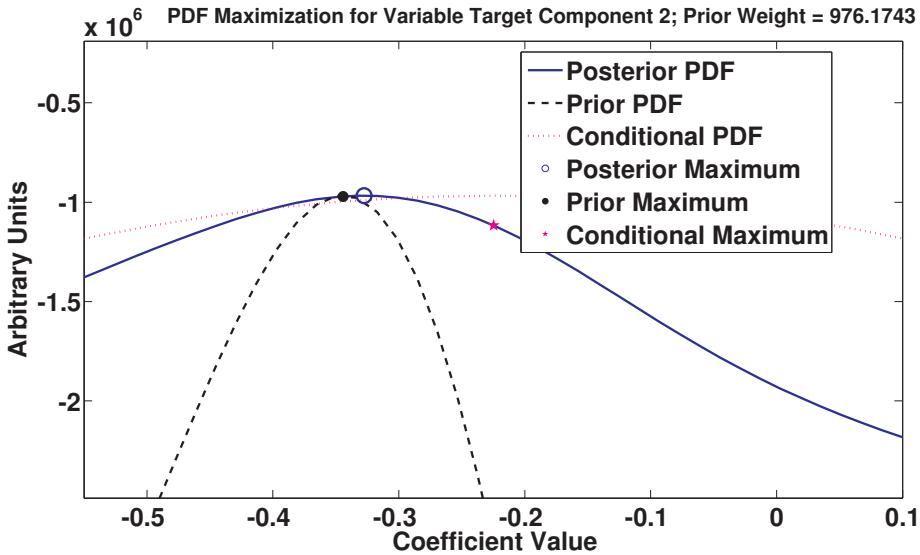
curve and so the conditional and posterior curves are virtually identical. Also, in cases where the width of the prior PDF is large, the maximum is low and makes little contribution to the posterior. A “wide” prior is equivalent to saying that the prior information results in little added information to a good fit, the extreme case being a “naive” prior PDF which provides no additional information. This is in contrast to the variable target where the conditional maximum is near the prior PDF, meaning the fit almost completely relies on the prior data, due to the high weight on the prior PDF of 976, resulting in only a slight influence of the conditional fit. In the case of the variable target in particular, if the prior PDF



**Figure 8.33** Posterior, prior and conditional PDFs for the Quiet Target. The prior and conditional curves have been renormalized to the same scale as the posterior for legibility. This target is quiet so the prior PDF weighting is low and does not influence the posterior much. The maximum of the posterior is therefore very close to the conditional maximum. The width of the prior PDF can also influence its height and amount of influence on the conditional. From Figure 18 of Smith et al. (2012).

was not determined using a weighted distribution (in  $K_p$ , RA, and Dec space) then the maximum of the prior would have been at about 0.6, as shown by the unweighted histogram in Figure 8.31. This would have resulted in a poor fit to the systematics. The actual prior fit takes into account the location of the TUS and the systematic trends in nearby targets.

Once the maximum of the posterior PDF is found for each basis vector the MAP fit is a linear combination of the basis vectors. The resultant fits are in Figure 8.35 and Figure 8.36. For the quiet target, all three fits roughly overlap the actual trend in the data. The prior fit is not an exact match and the slight disagreement is to be expected since the prior is purely formulated using targets other than the TUS. For a quiet target such as this one, highly weighting the prior PDF would result in a degradation of the fit, and so instead the PDF relies mainly on the conditional PDF (i.e. the red dashed and green solid curves over-

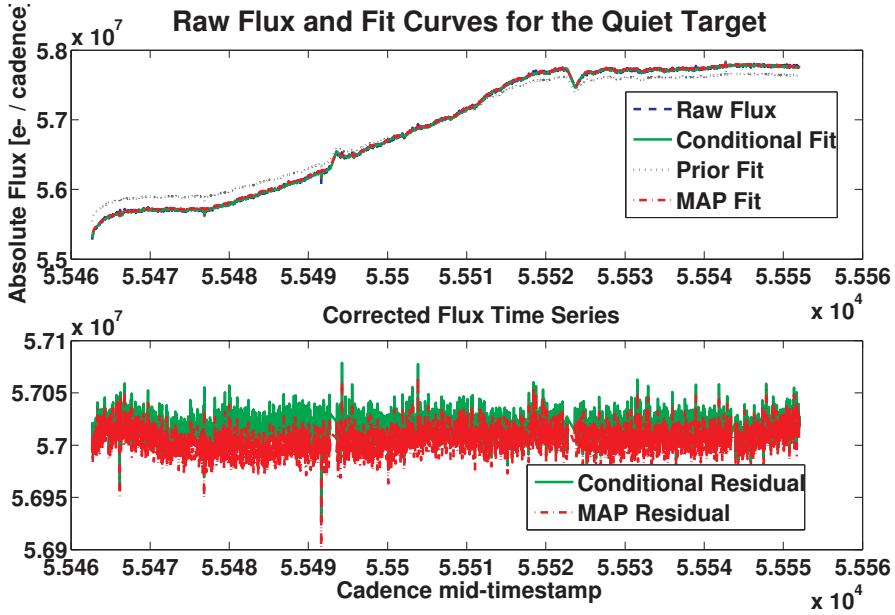


**Figure 8.34** Posterior, prior, and conditional PDF for the Variable Target. The prior and conditional curves have been renormalized to the same scale as the posterior for legibility. This target is highly variable so the prior PDFs is highly weighted at 976, which influences the posterior considerably. The maximum of the posterior is therefore close to the prior maximum. From Figure 19 of Smith et al. (2012).

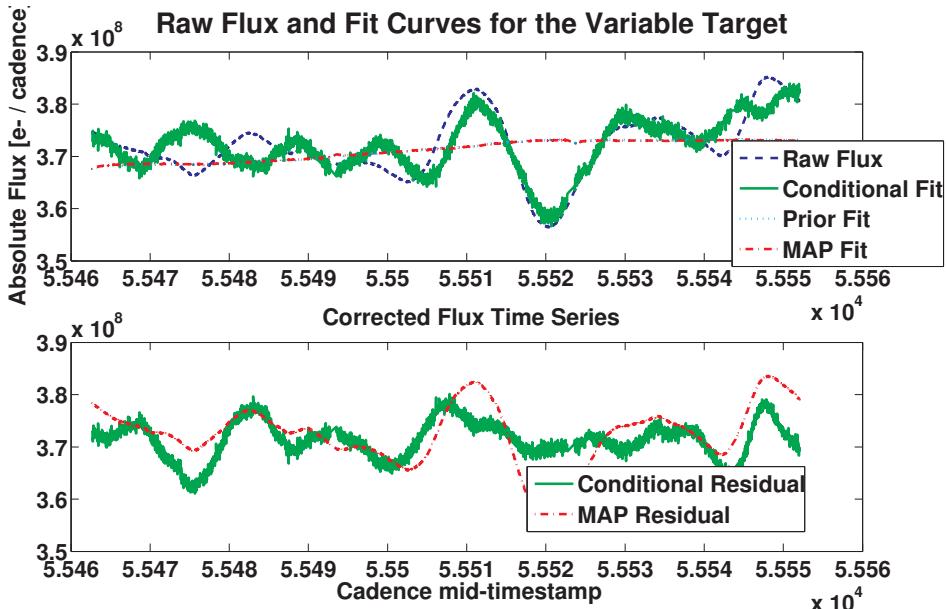
lap). The resultant light curve after the trend removal is in the bottom figure for both the MAP fit and the conditional fit, the latter being a LS fit. For the quiet target notice how the resultant curve is nearly featureless above the noise floor. Some slight artifacts are not fully removed and methods to correct these could be implemented. In the case of the variable target the conditional PDF results in a fit that attempts to remove all features in the light curve, whereas the prior PDF correctly identifies just the systematic trends in the data. In this case it is beneficial to rely principally on the prior PDF. The prior cannot be well discerned in the figure because it lies under the MAP fit. The conditional fit also introduces a considerable amount of noise into the corrected light curve due to its being constructed more from lesser basis vectors (shown in Figure 8.24) that contain larger noise components.

### 8.5.5 Iterating the Posterior with the Goodness Metric

The weighting on the prior is critical to the performance of MAP. The Prior Weight in Equation 8.30 has been found to not always be optimal. To better optimize the prior for each individual target, we iterate the posterior maximization with the goodness metric. In general, due to the clean nature of the prior, the greater the weight on the prior the lower the introduced noise. In contrast, the greater the weight on the prior the greater the bias and hence the more residual correlation due to a poorer LS fit to the trends. Therefore, we concentrate on the two goodness metric components: introduced noise,  $G_N$ , and residual correlation,  $G_C$ . We begin by maximizing the posterior with the prior weight as given in



**Figure 8.35** Resultant fits to the prior, conditional, and posterior (MAP) PDF for the Quiet Target. Target variability is low at 1.17 and  $W_{pr}$  is also low at 4.65. Quiet targets rely principally on the conditional PDF. From Figure 20 of Smith et al. (2012).



**Figure 8.36** Resultant fits to the prior, conditional and posterior (MAP) PDFs for the Variable Target. Target variability is high at 30.25 and  $W_{pr}$  is also high at 976.2. Variable targets rely principally on the prior PDF. It is evident that the Posterior (MAP) fit finds the systematic trend yet preserves the variability. From Figure 21 of Smith et al. (2012).

Equation 8.30. We then evaluate the goodness metric and scale the prior weight

using the following case structure:

$$\delta \mathbf{W}_{\text{pr}} = \begin{cases} \min \left( \frac{1}{20\Delta G_C}, 0.9 \right), & \text{if } G_C < 0.8; \\ \max (20\Delta G_N, 1.1), & \text{if } G_N < 0.8; \\ 1 & \text{otherwise,} \end{cases} \quad (8.31)$$

where

$$\Delta G_C = 0.8 - G_C \quad (8.32)$$

$$\Delta G_N = 0.8 - G_N \quad (8.33)$$

and poor correlation goodness takes precedence over poor introduced noise. The new prior weight will thus be:

$$\mathbf{W}'_{\text{pr}} = \mathbf{W}_{\text{pr}} \delta \mathbf{W}'_{\text{pr}}. \quad (8.34)$$

We then re-maximize the posterior and repeat the iteration until any of the following conditions is met:

- $G_C > 0.8$  and  $G_N > 0.8$ , or
- $\mathbf{W}_{\text{pr}} > 1.0 \times 10^7$ , or
- $\mathbf{W}_{\text{pr}} < 1.0 \times 10^{-3}$ , or stop if goodness is no longer improving. This speeds up the algorithm by not continuing to iterate with no improvements, or
- Iteration limit of 40.

### 8.5.6 Propagation of Uncertainties

Propagation of uncertainties is not necessarily straightforward because a covariance matrix is difficult to formulate for an empirical prior PDF. As a first approximation, the propagation can be assumed to be through a LS solution – which is close to the solution for most targets. If  $C_{\text{raw}}$  and  $C_{\text{cot}}$  denote the covariance matrices for the temporal samples of the raw and cotrended flux time series for a given target, then the uncertainties may be propagated (disregarding the uncertainty in the mean level which can be considered to be negligible) by:

$$C_{\text{cot}} = T_{\text{cot}} C_{\text{raw}} T_{\text{cot}}^T, \quad (8.35)$$

where the transformation  $T_{\text{cot}}$  is defined by

$$T_{\text{cot}} = (I - HH^T). \quad (8.36)$$

$H$  has the same design matrix as in Equation 8.19. This is overly conservative since the posterior PDF is more constrained than a simple LS fit. A more accurate propagation of uncertainties would take into account the attenuation of the uncertainties due to the prior PDF.

### 8.5.7 Application of PDC-MAP to Short Cadence Data

The principle issue with applying the MAP technique to SC data is the limited number of targets per module output. No more than 512 SC targets are collected at any time; these are spread over the entire FOV, so the number of SC targets per channel is small and at most about a dozen. A dozen is too small of a sample for the prior PDF or basis vectors to be properly formulated. Fortunately, all SC targets are also LC targets, so priors are already developed for them. A simple way to extend MAP to SC data is simply to use the basis vectors interpolated from LC data and the LC fit coefficients as the prior. This method is referred to as “quickMAP” since it utilizes a Bayesian-like PDF with a prior and a conditional PDF. A “Bayesian” formula can then be assembled as follows:

$$\theta_{sc} = (1 - \alpha)\theta_{MLE} + \alpha\theta_{lc}, \quad (8.37)$$

where  $\Theta_{lc}$  is the coefficient fit to the basis vectors from the LC PDC-MAP run (i.e. “the prior”),  $\theta_{MLE}$  is a LS fit to the SC time series of the interpolated LC basis vectors, and  $\alpha$  is a weighting parameter to determine to what extent to rely on the prior. As with normal MAP, the prior is weighted more for highly variable targets:

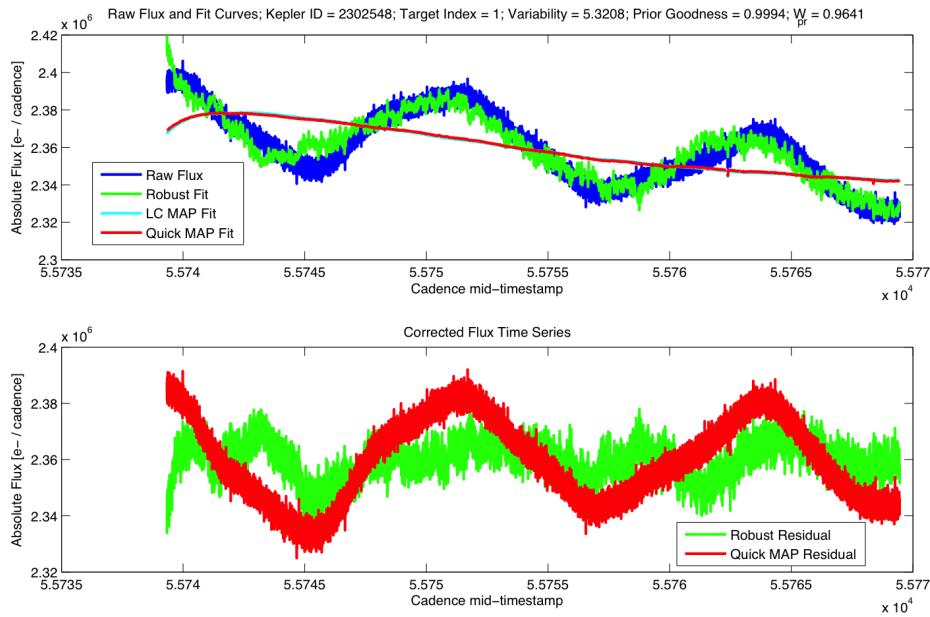
$$\alpha = 1 - \frac{1}{\max(1.0, \text{variability})^2}, \quad (8.38)$$

where variability is calculated for the SC data using the same method as described for LC data in Subsection 8.5.1, except here, we do not normalize by the median variability ( $\tilde{V}$ ). The greater the variability the more the prior is weighted – just as occurs in PDC-MAP.

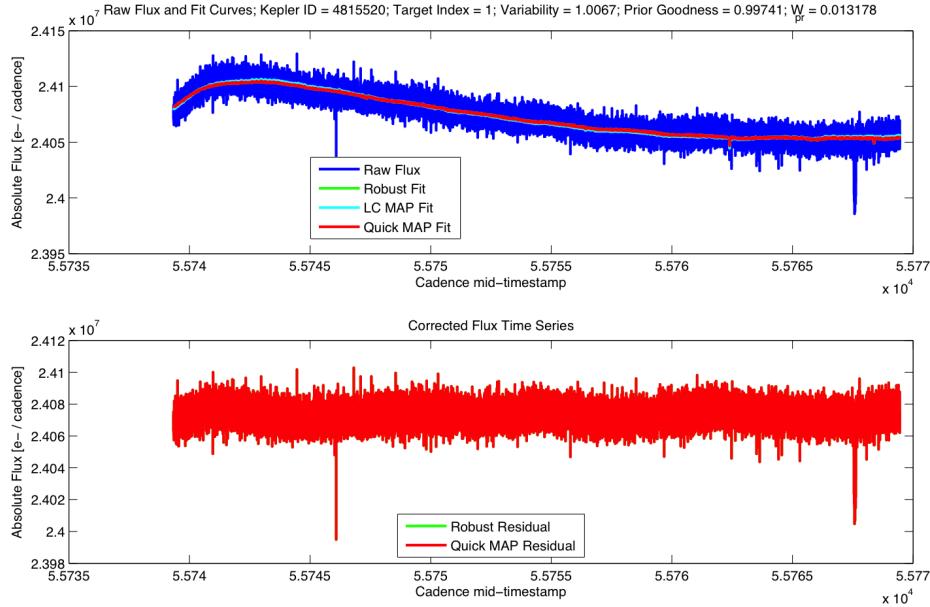
The basis vectors must be interpolated from LC to SC. The interpolation is performed with a simple spline function at a ratio of 30 to 1, which is high for a spline, but provides stable results. There is a potential concern that frequencies could be introduced above the LC *Nyquist* frequency; however, Fourier analysis has shown there to be virtually no introduced signals above this frequency. Provided that the LC data properly bracket the SC data with no gaps, the spline is well behaved. The interpolation means that no signals shorter than LC *Nyquist* frequency are corrected in SC data. Improvements to this method would be to include additional basis vectors that specifically address trends shorter than LC generated from the SC target data. However, with so few targets generating the trends, this will be difficult and prone to poor SNR.

For quiet targets  $\theta_{sc}$  relies principally on a simple robust LS fit to the interpolated LC basis vectors. For variable targets the LC fit is relied upon to control the posterior fit. For targets where the LC MAP fit was not performed,  $\alpha = 0$  and the fit relies solely on the robust LS fit.

Two examples of performance are given in Figure 8.37 and Figure 8.38.



**Figure 8.37** Example of a highly variable SC target. In this case, quickMAP principally relies on the LC MAP fit since  $W_{pr} = \alpha = 0.96$ .



**Figure 8.38** Example of a quiet SC target. In this case, quickMAP principally relies on the robust LC MAP fit since  $W_{pr} = \alpha = 0.013$ .

## 8.6 Multi-Scale MAP

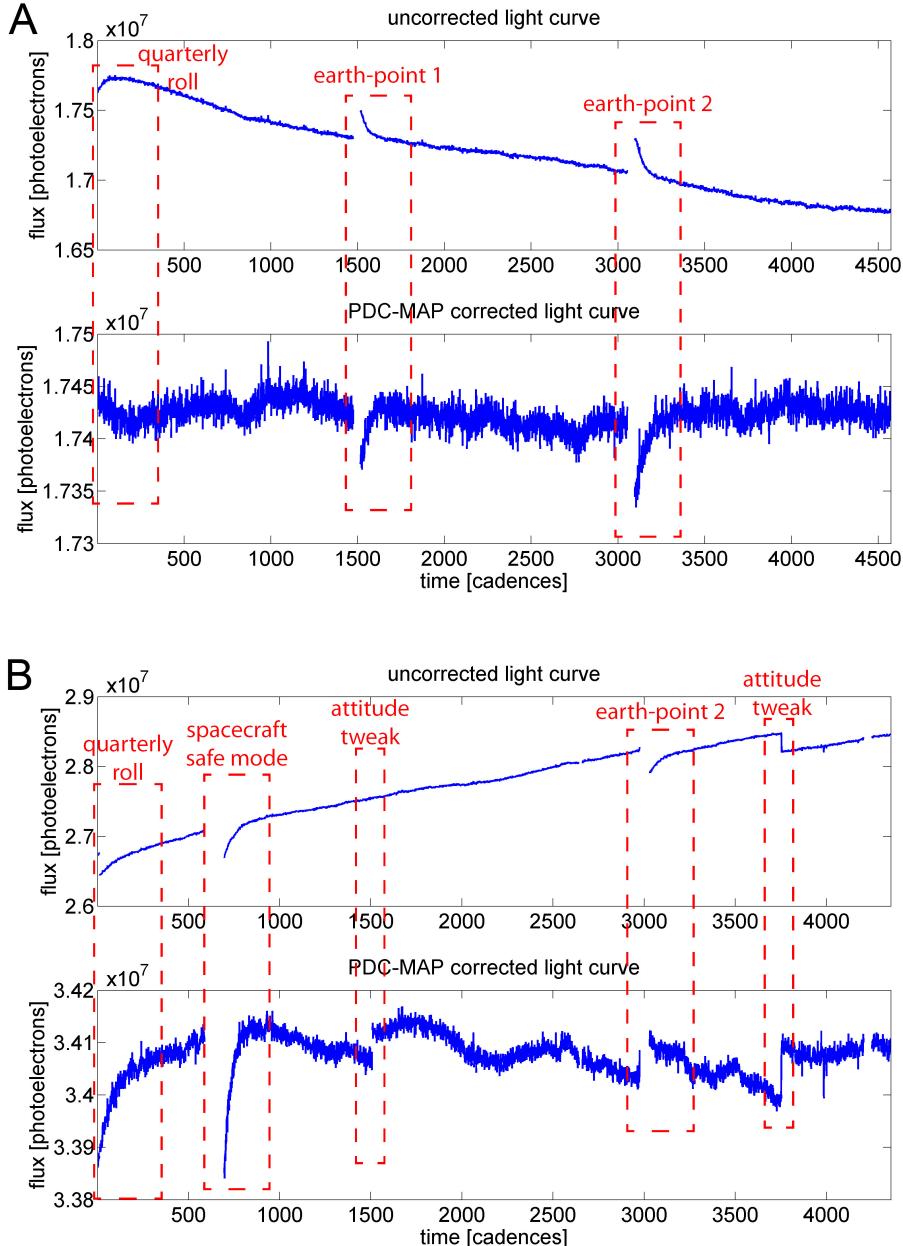
The PDC-MAP algorithm described in Section 8.4 performs very well for the majority of targets in the *Kepler* FOV. However, for an appreciable minority,

PDC-MAP has its limitations. To further minimize the number of targets where PDC-MAP fails to perform admirably, we developed a new method, called multi-scale MAP, or msMAP. Utilizing an over-complete discrete wavelet transform, the new method divides each light curve into multiple channels, or bands. The light curves in each band are then corrected separately, allowing for a better separation of characteristic signals and improved removal of the systematic errors.

### 8.6.1 Shortcomings of PDC-MAP

Correction of systematic errors in *Kepler* light curves has seen a dramatic improvement with the new Bayesian MAP detrending algorithm in the completely rewritten PDC module of *Kepler*. The main improvement of PDC-MAP over the previous PDC-LS method (Twicken et al., 2010a) is that the latter was prone to partially removing astrophysical signals and introducing significant noise for many processed light curves. The Bayesian MAP approach of the new algorithm makes it more robust against such overfitting. It can thus reliably remove the systematic errors while at the same time preserving the astrophysical features of the light curves. Despite this and other significant improvements to PDC, the corrected time series still exhibit some artifacts. First, about 20% of the light curves show some residual systematic errors. They also commonly exhibit incompletely corrected thermal transients from “Earth-point recoveries” (see Figure 8.39A). These are 100–200 cadence (2–4 days) long trends in the light curves that are caused by the thermal settling of the photometer after the monthly Earth-pointing events of the spacecraft or after its quarterly roll (Haas et al., 2010). A similar type of artifact can be seen in the recovery from safe mode events. For quarters that suffer from multiple interruptions of operation and commanded adjustments to photometer pointing, such as the highly challenging *Kepler* Q2 (June 2009 – September 2009), PDC-MAP will sometimes not perform error correction to a satisfactory degree (see Figure 8.39B). Second, for roughly 5%-10% of targets the systematic error correction introduces high-frequency noise, which can make detection of planet transits or analysis of astrophysical signals more difficult. The introduced noise is generally small but can be quite large for a handful of targets.

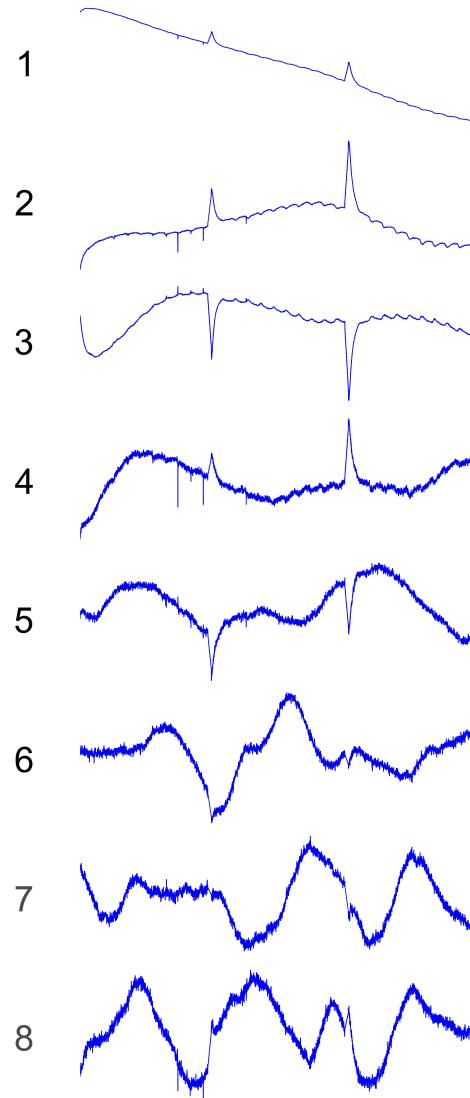
The above artifacts have the same origin: The MAP cotrending basis vectors, used to fit and remove the systematic errors in the light curves, usually contain features on very different timescales (see Figure 8.40), ranging from only a few cadences or hours (e.g., Argabrightenings, Witteborn et al., 2011), over several days (e.g. Earth-point recoveries, reaction-wheel zero crossings), to several weeks or even months (e.g. long trends due to DVA or focus changes). Thus, corrections of errors on different scales cannot be independent, and so the removal of errors on one scale can have the side-effect of injection or incomplete removal of errors on another scale. As a second issue, some basis vectors contain very high-frequency components and noise. This leads to injection of high-frequency noise in the MAP correction. In mathematical terms, these issues can be re-



**Figure 8.39** Worst case scenario examples of incomplete systematic error removal. A) A common case where Earth-point recoveries, and to a lesser extent the quarterly roll recovery, are not completely corrected. B) A light curve from Q2 with multiple imperfect corrections of safe modes, Earth-points, quarterly-rolls, attitude-tweaks, and loss of fine-point. From Figure 1 of Stumpe et al. (2014).

garded as a consequence of the set of cotrending basis vectors not forming an independent basis. The basis set is quite complete in the sense that all the trends are represented in the basis vector set. But because they are convolved with each other, proper removal is not always possible. Simply increasing the number of

cotrending basis vectors in the fit does not only quickly render it computationally infeasible, but further has not been found to significantly increase the overall performance of the systematic error removal (see Subsubsection 8.6.3.4). If anything, increasing the number of basis vectors just results in more stellar features being removed.



**Figure 8.40** Example of a set of eight non-multi-scale MAP Cotrending Basis Vectors (shown here for module output 7.3, Q10) exhibiting the issues motivating the development of multi-scale MAP. Systematics are present in the same basis vector on different timescale as is high-frequency noise for the latter basis vectors. From Figure 2 of Stumpe et al. (2014).

We note that, since the Earth-point recoveries are the most obvious type of residual systematic errors in PDC-MAP, other approaches to mitigate their detrimental effect on the light curve quality are conceivable. One approach would be

to simply ignore the recovery phase and discard the respective cadences. However, this would discard a substantial portion (about 10%) of all data points in each time series. It is important to realize that this loss could not be mitigated by longer *Kepler* observation times because the recurring gap in the data would create a blind spot for planet transits with a corresponding epoch and period to fit into these gaps. Another conceivable approach could be to change the unit of work in *Kepler* PDC from quarter-long light curves to only month-long light curves. This would yield only one strong thermal transient signal at the beginning of the time series which might be easier to correct. However, this approach would limit the maximum length of systematic errors that PDC can identify and correct to one month because the maximum length scales of stellar features that can be preserved are limited by the length of the unit of work. Further, this approach would only work for recoveries from scheduled events (such as the monthly Earth-point), but not with recoveries from unplanned events such as safe modes or loss of fine point (e.g. Q2), which lead to similar recovery artifacts. To avoid these and other potential shortcomings of simpler approaches, we have designed a more sophisticated algorithm that sacrifices neither data quality nor quantity.

### 8.6.2 The Solution: Multiscale Error Correction

In this section we present a solution to these problems, which we have first implemented as a major improvement to *Kepler* PDC in version 8.2. The approach we take is to perform a separation of scales in the time series, such that small scale features and large scale features are described by different cotrending basis vectors. Figure 8.41 illustrates this approach: Each time series is split into multiple channels (bands). The set of light curves in each band is corrected separately with the MAP error correction algorithm. The corrected light curve bands are then combined again to generate the corrected light curves. Since many different systematics occur on different timescales this *band splitting* is useful in isolating the systematics. For example, thermal stresses on the spacecraft due to its orbit about the sun results in systematics on a yearly timescale, whereas the reaction wheel heater cycling occurs on a three-day cycle. These two systematic effects are independent of each other and since they occur on different timescales they can be isolated using the band splitting method described below.

To decompose each light curve into a set of light curves on dyadic (power of two) scales, we use an overcomplete discrete wavelet transform (Jenkins et al., 2002). As a joint time-frequency representation, it is a natural choice for such a multiscale analysis. Similar to the short-time Fourier transform (STFT), the wavelet transform is a windowed analysis technique that allows for the analysis of non-stationary signals. In contrast to the STFT, however, the wavelet transform uses variable-sized windows to tile the time-frequency plane. In particular, the wavelet transform employs basis functions  $\Psi_{\tau,\lambda}(t)$  that are scaled (by  $\lambda$ ) and shifted (by  $\tau$ ) versions of a mother wavelet  $\Psi(t)$ , and that are finite and

vary in both duration and bandwidth. Because of these properties, the frequency and temporal resolution vary across scales in such a way that their product (the area of the time-frequency tile) is constant at all scales. This property, that the bandwidth at each channel divided by its center frequency is constant, is also known as the *Constant Quality* (“Constant-Q”) property (Vetterli & Kovacevic, 1995). As a result, the wavelet decomposition achieves the best time resolution at the shortest scales (highest frequencies) and the best frequency resolution at the longest scales (lowest frequencies).

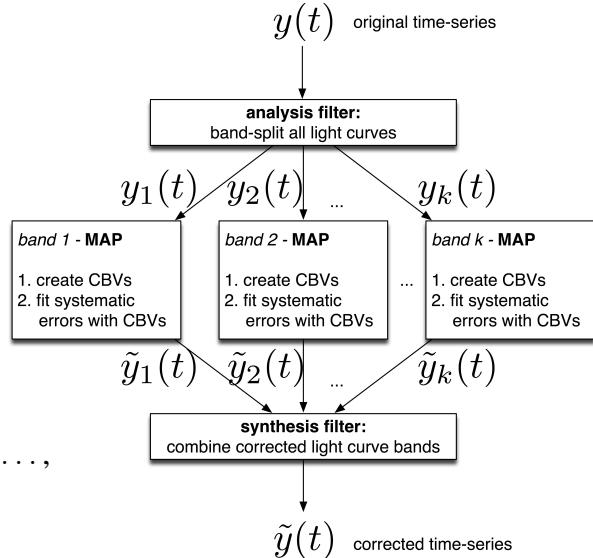
Convolution of a signal  $y(t)$  with the wavelet basis functions  $\Psi_{\tau,\lambda}(t)$  via the *Wavelet Transform*,  $WT$ , produces a series of wavelet coefficients  $w(\tau, \lambda)$  for each wavelet basis function (characterized by its shift  $\tau$  and scale  $\lambda$ ),

$$w(\tau, \lambda_i) = WT[y(t)]. \quad (8.39)$$

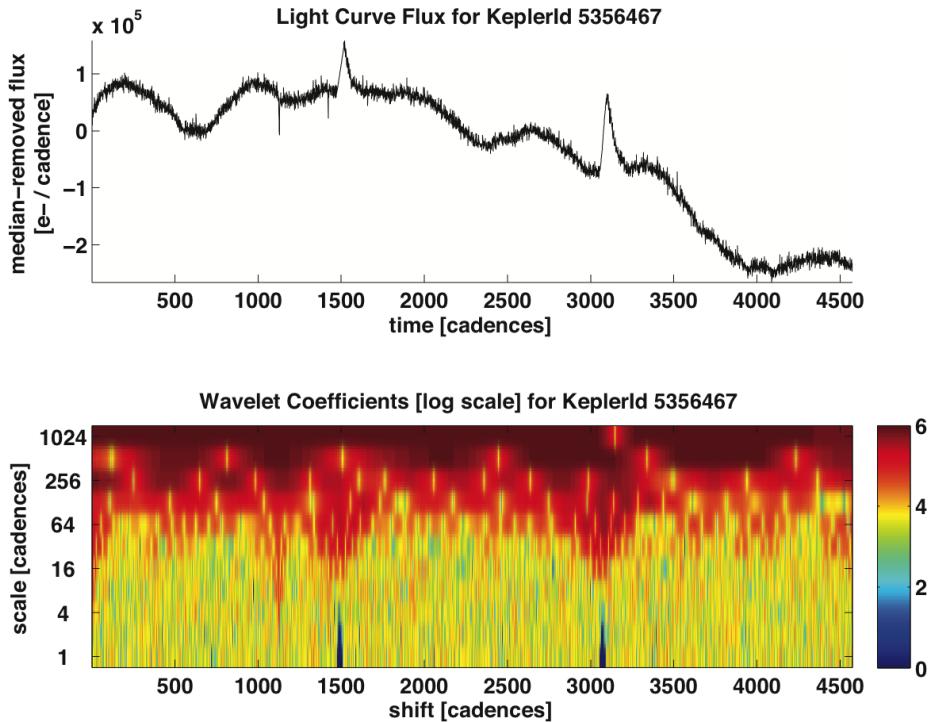
The scales are chosen as powers of two ( $\lambda_i = 0, 1, 2, 4, 8, \dots, N$ ), leading to a doubling of the characteristic scale in each band and the constant-Q property.

By taking all possible shifts in each band ( $\tau_j = 1 \dots T$ , where  $T$  is the length of the discrete input time series  $y(t)$ ), we perform an overcomplete discrete wavelet transform (Jenkins, 2002). Figure 8.42 shows the original light curve and the set of wavelet coefficients for the light curve. The wavelet coefficients  $w(\tau, \lambda)$  are proportional to the power of the signal at each particular shift  $\tau$  and scale  $\lambda$ . Here we can see that distinct events, such as the Earth-point thermal recoveries, are spread out as we move to longer length scales. Long-term trends are in the longest scale (1,024 cadences); however, the colormap has been saturated at large values in order to show the details at smaller scales. Intermediate length features appear in the 128, 256, and 512 cadence scales. High-frequency features are in the 1–4 cadence scales. Also note that the single spike in the light curve at cadence 1,150 exhibits itself as a streak along many scales.

Wavelet coefficients are not used for MAP directly, but instead we immediately perform an inverse wavelet transform in each band, reconstructing a band-



**Figure 8.41** The Multiscale PDC-MAP correction scheme. Each light curve  $y(t)$  is decomposed into  $k$  bands  $y_i(t)$  ( $i = 1 \dots k$ ) using an analysis filter. The MAP systematic error correction algorithm is performed on each band  $y_i(t)$  separately to generate the corrected light curve band  $\hat{y}_i(t)$ . Finally, the corrected light curve bands are combined again with a synthesis filter, yielding the corrected light curve  $\hat{y}(t)$ . From Figure 3 of Stumpe et al. (2014).



**Figure 8.42** Wavelet analysis of the light curve for KIC 5356467 Q10. The top panel shows the input light curve flux. The lower panel shows the wavelet coefficients  $w(\tau, \lambda)$ , where  $\tau$  is the shift, plotted horizontally, and the scale, which is  $2^{\lambda-1}$ , on the vertical axis. Distinct features in the light curve can clearly be seen extending across multiple scales in the wavelet coefficients. In order to show details at shorter scales, the 1024 scale is saturated in the colormap. From Figure 4 of Stumpe et al. (2014).

split light curve in the time domain. It is this band-split signal in the time domain to which MAP is applied. With the overcomplete wavelet transform the reconstruction of the whole signal  $y(t)$  from the wavelet coefficients  $w(\tau, \lambda)$  can be done by applying the inverse transform  $WT^{-1}$  in each band  $i$  separately:

$$y_i(t) = WT^{-1} [w(\tau, \lambda_i)]. \quad (8.40)$$

Taking the linear sum over  $y_i(t)$  for all scales  $i$  would result in the original light curve before band splitting:

$$y(t) = \sum_{i=1}^N y_i(t). \quad (8.41)$$

But because this reconstruction and the MAP correction are both linear operations, we can perform a MAP correction in each band separately:

$$\tilde{y}_i(t) = \text{MAP} [y_i(t)] \quad (8.42)$$

and then take the sum of the corrected light curve bands,  $\tilde{y}_i(t)$ , to obtain the total corrected light curve  $\tilde{y}(t)$ :

$$\tilde{y}(t) = \sum_{i=1}^N \tilde{y}_i(t). \quad (8.43)$$

Note that even though the MAP operation is linear it is not commutative:

$$\sum_{i=1}^N \text{MAP}[y_i(t)] \neq \text{MAP}\left[\sum_{i=1}^N y_i(t)\right], \quad (8.44)$$

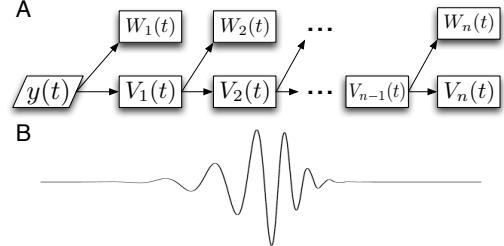
and so the band-split msMAP operation results in a distinct correction to regular MAP.

The process of a wavelet transformation followed by a inverse wavelet transformation can be interpreted as an octave filterbank that iteratively splits a signal  $V_i(t)$ , with the input signal  $y(t) = V_0(t)$  into a ‘detail’ layer  $W_{i+1}(t)$  and an ‘average’ layer  $V_{i+1}(t)$  as illustrated in Figure 8.43A. The  $i^{\text{th}}$  “detail” layer captures changes in the input signal on a scale of  $2^{i-1}$ , while the  $i^{\text{th}}$  ‘average’ layer captures the smoothed structure on a scale of  $2^i$ . The perfect reconstruction property of the wavelet decomposition guarantees that the original signal is the sum of the ‘detail’ layers and the last “average” layer:

$$y(t) = V_n(t) + \sum_{i=1}^N W_i(t). \quad (8.45)$$

In light of Equation 8.45, we can now refer back to Figure 8.42 to discover that the plotted wavelet coefficient scales 1–512 are the ‘detail’ layers  $W_i(t)$ , whereas the 1024 scale is actually the final ‘average’ layer  $V_n(t)$ .

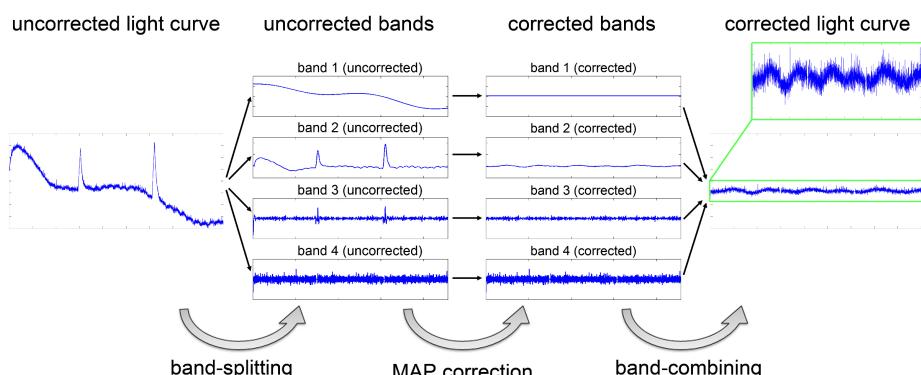
As the mother wavelet, we use the Daubechies 12-tap wavelet (Cohen et al., 1992) (see Figure 8.43B). It is important to note that the MAP fit is not performed in the wavelet domain, but in the regular time domain. The process is illustrated in Figure 8.6.2. A further detail not shown in the figure is that the uncertainties are split using the same method as on the light curve data and then the propagation of uncertainties method as used in MAP is applied in each band. The uncertainties are then combined to produce the output uncertainties for msMAP.



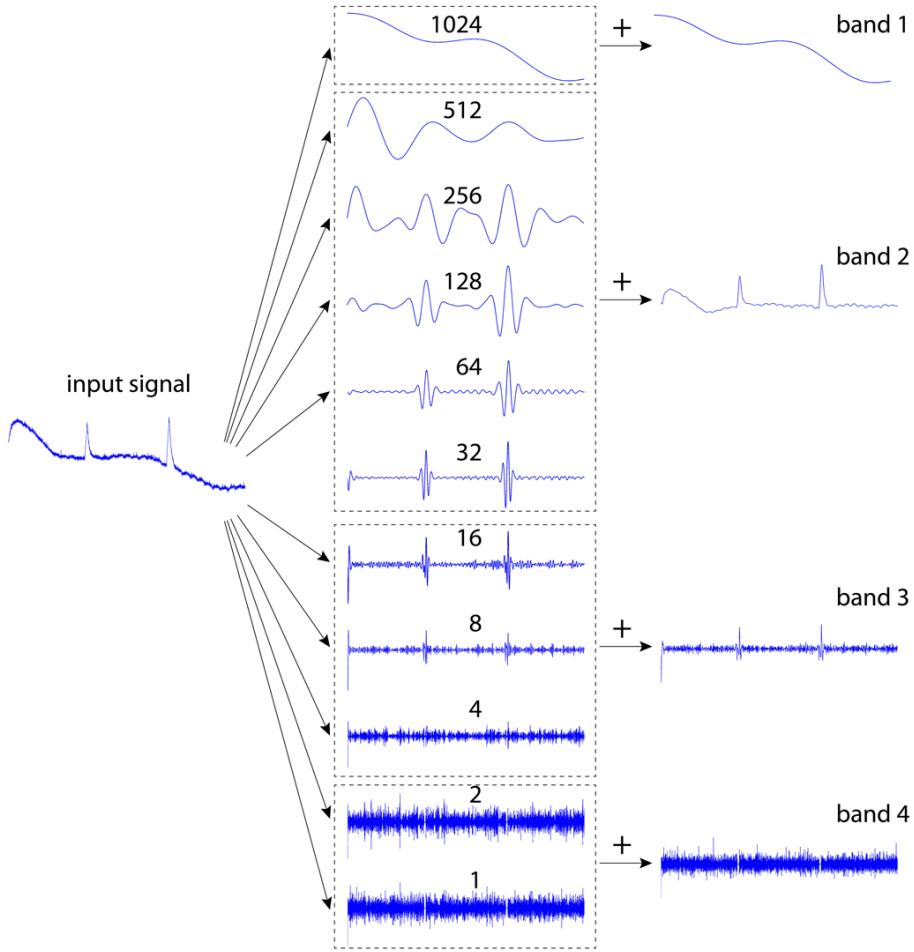
**Figure 8.43** Illustration of the wavelet transformation. A) Represented as an octave filterbank which decomposes the input signal into successive ‘detail’ layers  $W_i(t)$  and ‘average’ layers  $V_i(t)$ . B) Daubechies 12-tap wavelet, which is chosen as the mother wavelet in this work. From Figure 5 of Stumpe et al. (2014).

A wavelet decomposition of a typical light curve into its channels (1–11) is shown in Figure 8.45. Each channel has a characteristic scale that is twice as large as that of the next channel. Rather than performing a MAP correction directly on each individual channel, however, it is preferable to group a number of adjacent channels together and perform the MAP correction on each of these combined bands. In the case illustrated in Figure 8.45, for instance, 11 channels are combined into 4 bands. Aside from the MAP correction being relatively expensive to perform on 11 channels individually, we found that the correction performance is better when grouping channels together this way. The main reason for this is that when having many bands, the errors and features in the time series are spread across band boundaries and are distributed across multiple bands. Thus, different parts of the same feature are subjected to different MAP fits, which can lead to imperfect corrections. We therefore choose the groupings so that characteristic features are wholly contained in a single band. Also, overfitting can occur if all 11 channels are fit separately, due to the large number of degrees of freedom in the correction (see Subsubsection 8.6.3.4).

Figure 8.46 gives the first four MAP cotrending basis vectors in a three-band decomposition. The cotrending basis vectors in the different bands separate the scales of the systematic errors very neatly, allowing independent correction of the errors. In particular, band 1 contains long-term trends, band 2 contains artifacts of medium duration (such as the Earth-point recoveries and the three-day reaction wheel cycle), and band 3 contains very high-frequency features such as Argabrightenings and high frequency oscillations. For band 3, which contains the characteristic scales of 1 and 2 cadences, the SNR test discussed in Subsection 8.5.1 finds no significant basis vectors, due to the noise floor being approximated by the first differences in the flux. For a set of vectors with signals only up to 2 cadences long, there is not enough of a distinction between the first differences and the true signal for the SNR test. Instead we rely on Bayesian Model Selection (Minka, 2008) to find the proper dimensionality in Band 3. Typically only one or two basis vectors are found that contain high-frequency oscillations. When MAP finds no significant systematic signals among the light curves, it ve-



**Figure 8.44** Bandsplitting example. From Figure 6 of Stumpe et al. (2014).

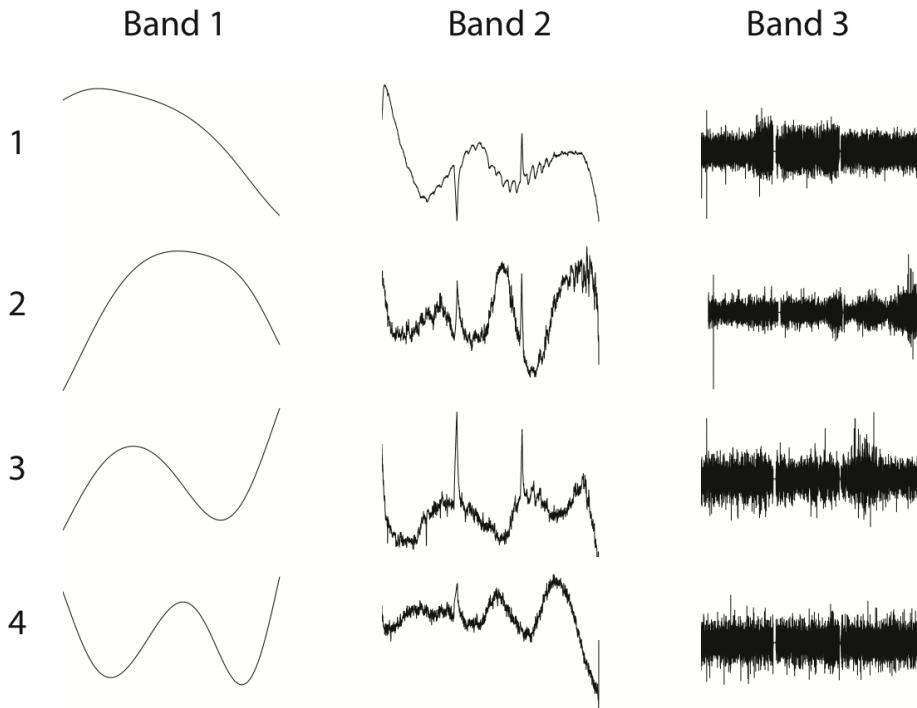


**Figure 8.45** Decomposition of the input light curve (left) into a set of sub-bands with dyadic scale. Multiple sub-bands are combined (dashed boxes) by summation to yield the final bands (right). The numbers on the sub-bands denote the characteristic scale in units of cadences for the respective sub-band. From Figure 7 of Stumpe et al. (2014).

toes all cotrending basis vectors. With no surviving cotrending basis vectors, no correction is performed and the light curve signals are simply preserved. This aids in separating the noise from the systematic errors.

### 8.6.3 Choice of Parameters

The PDC-MAP algorithm has a set of parameters, such as the number of basis vectors and certain weighting coefficients that can be chosen by the pipeline operator to achieve optimal correction performance. The multiscale extension presented here introduces more parameters and detail choices for the algorithm that have to be characterized and optimized for correction performance. In particular, the most important parameter choice is the grouping of channels to combine into bands, which determines the number of bands as well as the scale range of



**Figure 8.46** The first four MAP cotrending basis vectors in a three-band decomposition for Q5 Module Output 7.3. Signals are principally in the first two bands but there are slight signals in band 3. From Figure 8 of Stumpe et al. (2014).

each band. Moreover, the PDC-MAP parameters can be chosen independently for each band, leading to a significantly larger parameter space. We have tested a wide range of parameter choices to investigate their effects on the correction performance and we discuss our main findings here.<sup>15</sup>

**8.6.3.1 Number of Bands** The number of bands directly determines the number of degrees of freedom for the overall MAP correction, and thus the “flexibility” of the fit. The two border cases are: 1) performing a MAP fit on each channel without any grouping of bands and 2) grouping all channels together into one band, equivalent to the regular version of PDC-MAP. We found that the best results are usually achieved by using a decomposition into either three or four bands. Fewer bands lead to residual artifacts as in the case of regular PDC-MAP, whereas more than five bands can lead to unconstrained fits, splitting of features across multiple bands, and either overfitting or introduction of artifacts in extreme cases. A general trend we observed is that time series suffering from more and stronger artifacts (e.g. heavily corrupted observation quarters such as Q2, more sensitive CCD channels) benefit from corrections with four or even five bands, whereas for time series with fewer systematic errors, three bands are

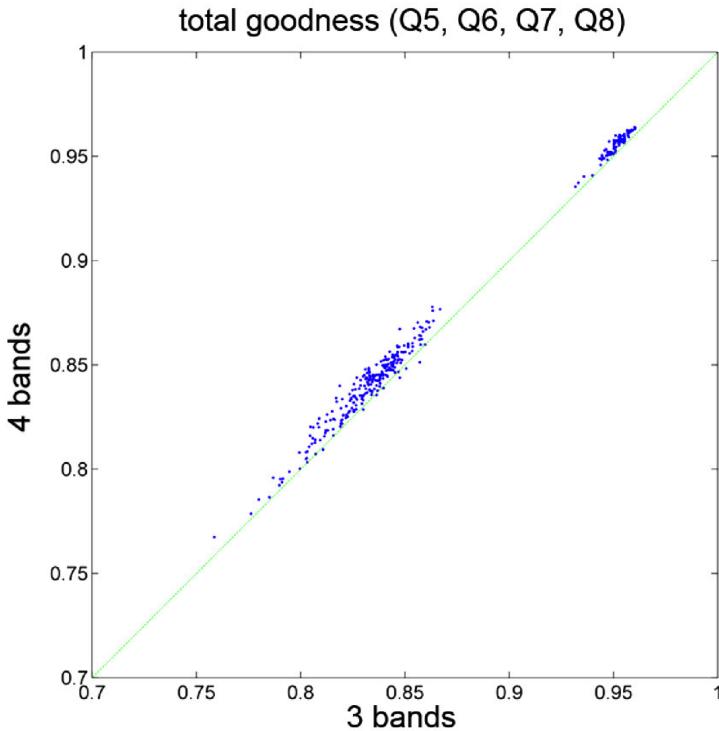
<sup>15</sup>It should be noted that the current *Kepler* Pipeline architecture does not allow for specific parameters for individual channels but only for individual quarters where all channels in each quarter use the same parameters.

usually the best choice. To perform a comprehensive comparison between 3- vs 4-band decomposition, we performed both on all CCD channels of quarters Q5–Q8, (i.e. a time span of one full year to exclude potential seasonal effects) and evaluated the correction quality using visual investigation of a subset of light curves, as well as with the PDC goodness metric to obtain aggregate statistics (see Subsubsection XIV for details of the goodness metric).

The PDC goodness metric quantifies the performance of the cotrending in PDC with four basic performance qualities: 1) removal of target-to-target correlations, 2) injection of noise, 3) preservation of stellar signals, and 4) removal of Earth-point thermal recoveries. The study concluded that the absolute magnitude of the differences in the goodness metric between the 3- and 4-band decomposition is only marginal and is almost negligible in practice. This is evident when comparing the total goodness between 3 and 4 bands as shown in Figure 8.47. This shows that for almost all CCD channels the difference between 3-band and 4-band average correction performance is on the order of only 1%. In the visual comparison of the corrections for individual targets, we did find some examples where the correction was better using four bands (see Figure 8.48A for an example), and in rare cases the 4-band correction was significantly worse than the 3-band correction (see Figure 8.48B for an example). The bottom line of this comparison is that either three or four bands should be used and they both yield very similar correction performance.

**8.6.3.2 Channel Grouping, Band Boundaries** The second central parameter choice involves selecting which channels – and thus characteristic scales – to group together and treat in the same MAP fit. For example, Figure 8.45 shows a decomposition where band 1 contains only the scale of 1024 cadences, band 2 contains the scales 512, 256, 128, 64, 32; band 3 contains the scales 16, 8, 4; and band 4 contains the scales 2, 1. We abbreviate this here as “(1024 / 512,256,128,64,32 / 16,8,4 / 2,1)”. Alternative 4-band decompositions could be (1024 / 512,256,128 / 64,32,16,8,4 / 2,1) or (1024,512,256 / 128,64,32 / 16,8 / 4,2,1). We systematically varied the band boundaries for 3- and 4-band decompositions, and identified two important factors.

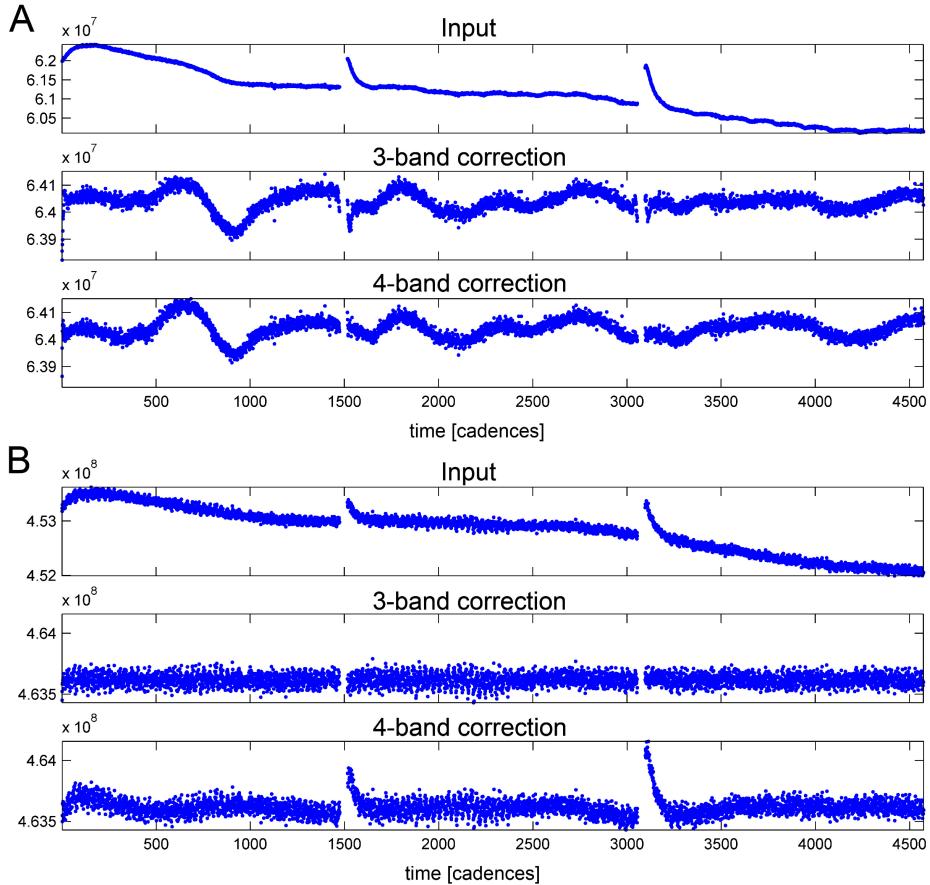
First, the most important separation is between 1) very long trends on the scale of weeks or months ( $> 800$  cadences) and 2) the three-day ( $\sim 150$  cadences) Earth-point recoveries and the  $\sim 3$ -day reaction wheel desaturation cycles (see Van Cleve & Caldwell, 2016). It is therefore favorable to not have the 1024 channel in the same band as the 256 or 128 channel. Second, it is beneficial to have a separate band for very short scale features, such as the (2,1) band. This helps to decouple the noise from the correction of intermediate-scale or large-scale errors and helps to reduce the problem of noise injection in PDC-MAP significantly. We also found that having most of the spectrum of a particular systematic error in the same band helps to improve the correction quality. For instance, Earth-point recoveries have a strong signal on the scales (128,64,32) (see Figure 8.42). Since Earth-point recoveries are the most prominent residual



**Figure 8.47** Comparison of the total goodness between a 3-band vs. a 4-band correction for all four quarters Q5–Q8. From Figure 9 of Stumpe et al. (2014).

systematic errors in PDC-MAP, grouping the channels with scales (128,64,32), and possibly also 256, in the same band appears useful for the correction. Based on these observations, we found (1024 / 512,256,128,64,32 / 16,8,4 / 2,1) to be a good 4-band decomposition and (1024 / 512,256,128,64,32,16,8,4 / 2,1) to be a good 3-band decomposition for most cases. However, as with the number of bands, the overall correction performance was not overly sensitive to this parameter choice within small variations.

**8.6.3.3 Modification of MAP Parameters** An investigation of several MAP parameters, in particular the number of basis vectors per band, showed that most MAP parameters should remain unchanged and the same for each band. One parameter that had to be changed is the light curve normalization method for the SVD and the MAP fit. We use normalization by the mean only in the longest-scale band, because the other bands have zero mean. Instead, the medium-scale bands are normalized by the standard deviation of each light curve. With no long-term trends (or high-frequency noise) in the medium scale bands the standard deviation remains the best metric for normalization. The shortest scale band is normalized by the noise floor which is estimated by the first differences between the cadence flux values. Another MAP parameter that must be tuned is the prior PDF goodness weighting component in the prior weighting calculation. The gain



**Figure 8.48** Comparison of a 3-band vs a 4-band correction. A: a 4-band correction can sometimes lead to better results when the 3-band correction has residual systematic errors. B: on the other hand, 4-band corrections can sometimes introduce artifacts. Both of these cases are rare, and usually both corrections perform similarly well. From Figure 10 of Stumpe et al. (2014).

in the prior PDF goodness must be increased for the shorter bands relative to the stellar variability component of the prior weight.

A final noteworthy parameter change is that we perform a robust LS fit, rather than a full MAP fit, in the longest scale band. We found that a MAP fit in the longest scale band can lead to artifacts in the form of low-frequency waves. These waves, which are usually very small, are exacerbated by non-ideal MAP fit priors in this band. In the case of bad priors, MAP usually sets the weight of the prior to zero, effectively reverting back to a robust LS fit. This successfully reduces the artifact. However, in the case of the longest scale band, even a slight error on the prior can bias the posterior fit too far away from the conditional resulting in an artifact residual wave in the light curve. The robust LS fit does introduce the real risk of overfitting but we have found that the additional attenuation of long period signals in msMAP is small compared to regular MAP and acceptable given the much greater performance at shorter frequencies relevant to transit detection. Therefore the default configuration is to perform a robust LS fit

in the longest band. It is difficult to distinguish between systematic and intrinsic stellar signals at periods approaching one observing quarter in length and so we believe the forced robust fit is acceptable, but we are investigating methods to improve the prior in the longest scale band and preserve signals out to longer periods.

**8.6.3.4 Discussion of Alternative Approaches** In addition to the optimization of parameters, we have also investigated alternative approaches and modifications to the algorithm. We will briefly describe the most relevant ones here.

- **Increasing the number of basis vectors without employing a multiscale framework**

Because the systematic errors we want to correct are on a variety of different timescales, a multiscale approach seems very natural. However, it is an interesting question whether the dramatic performance improvements achieved by our new algorithm are really due to the multiscale aspect of our approach. One side effect of the multiscale approach is that it effectively increases the total number of basis vectors, and thus the degrees of freedom in the MAP fit. Could similar performance be achieved by simply increasing the number of basis vectors in a regular MAP fit? We have tested the effect of the number of basis vectors in the original PDC-MAP work and found that going beyond eight basis vectors does usually not improve the correction performance. PDC-MAP chooses the optimal number of basis vectors automatically based on the eigenvalue spectrum, and tests where we explicitly forced the number of basis vectors to be larger (e.g. 16, 24) did in fact not show any performance improvement.

- **Using multiscale basis vectors in a joint fit**

Another conceivable option, and in fact an alternative design that we tried initially would be to generate the basis vectors separately for each band – as is done here – but instead of performing one MAP fit in each band separately, just do one MAP-fit with the joint set of basis vectors on the original unsplit light curves. We have tried this approach, but found that it does not perform well. Almost all light curves showed strong residual systematic errors or even injection of artifacts, such as enhancement of the Earth-point thermal transients. This is due to overfitting since the number of basis vectors used is expanded by a factor of three. Thus, the basis vectors for each band should be fit separately on band-split light curves. Essentially, this approach suffers from similar problems as the original PDC-LS (Twicken et al., 2010a), the predecessor of PDC-MAP, and these deficiencies were the initial motivation to develop the latter.

- **Processing of all channels without grouping them into bands**

One central step in our algorithm is the grouping of several adjacent octave channels into bands (see Figure 8.45). This additional step introduces ad-

ditional parameters (i.e. the band composition, see Subsubsection 8.6.3.2) and it is questionable whether a simpler design without this step would work equally well. We tested this and performed a msMAP correction on all individual channels, but found that the results are significantly worse. In particular, using such a large number of bands renders each individual fit too unconstrained and leads to substantial overfitting and removal of many astrophysical signals. In some rare cases where Earth-point artifacts were not fully corrected with a 3-band or 4-band fit, this approach performed better at removing those artifacts at the expense of severe overfitting.

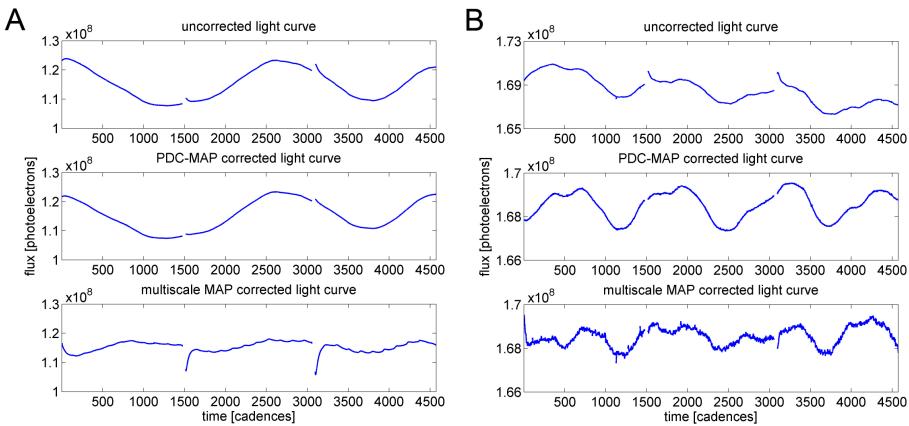
- **Choice of the wavelet family for band-splitting, and alternative filter-banks**

Daubechies-wavelets have several useful characteristics that make them a common choice for a mother wavelet. One of their major advantages is that an orthogonal set of Daubechies-wavelets can be created (Cohen et al., 1992), which renders the synthesis process into a simple summation of the individual bands – a fact that we exploit (Eq. Equation 8.41 and Equation 8.43). Two other commonly used wavelet families are the simple Haar-wavelets, and the Gabor-wavelets. Both are widely used in image processing and computer vision, for instance in feature-based object detection (Zeng et al., 2009). Gabor-wavelets have the disadvantage that they are not orthonormal and thus the signal synthesis is a more complex operation. While this does not play a role in its intended application where synthesis is usually not required, it would complicate processing in our case application here. Therefore, we have only investigated using Haar-wavelets as alternative to Daubechies-wavelets. The correction performance using Haar-wavelets was similar but inferior in most cases, most likely due to it having poor frequency response, resulting in large leakage out of each band. Finally, simpler approaches for constructing the filterbank for band-splitting are also conceivable. In particular, we tested band-splitting using mean-filters, median-filters, Gaussian-filters, and Savitzky-Golay filters (Savitzky & Golay, 1964), but we found that the overall correction performance was best when using wavelets. This result is not surprising, given that the scale-invariance of wavelets makes them intrinsically suited for multiscale-related problems and accounts for their great success in these applications.

#### 8.6.4 Further Algorithm Details

**8.6.4.1 Cases of Bad Corrections: Vetting** In some cases (1–2% of all targets) it can happen that msMAP fails in correcting a light curve and that systematic errors or noise are enhanced. Fortunately, these bad corrections are usually not subtle imperfections but rather notable failures that are obvious upon inspection. In these cases, the more conservative correction of the original PDC-MAP usually yields a better correction. Figure 8.49 shows two representative examples

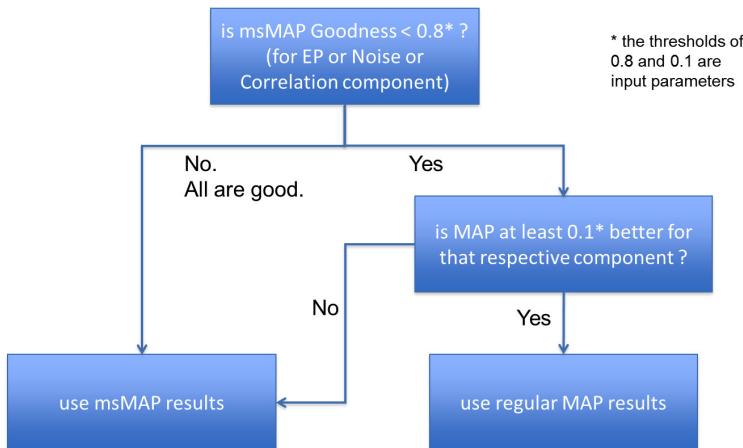
of such cases. In the first case (Figure 8.49A), the msMAP correction (lower panel) substantially enhances the Earth-point recovery artifacts and the three-day reaction wheel cycle. In contrast, the regular PDC-MAP correction (middle panel) does not correct the Earth-points recoveries completely, but the overall correction quality is much better. In the second case (Figure 8.49B), the msMAP correction shows significant injection of high-frequency noise, and also exhibits overfitting. The regular PDC-MAP correction, in contrast, appears nearly flawless. The main cause for these bad corrections are bad priors for the MAP fit, mostly in the middle bands. Many of the targets with bad corrections are very bright ( $K_p \leq 10$ ). Since the magnitude has a strong influence on the MAP prior, the sparsity of correlated targets at low magnitude used to generate the prior can explain the bad priors for many of these targets.



**Figure 8.49** Two cases of bad corrections with msMAP, showing strong residual artifacts such as Earth-point recoveries and three-day reaction wheel cycles (panel A), as well as noise-injection and overfitting (panel B). From Figure 11 of Stumpe et al. (2014).

PDC tries to identify these bad corrections using the goodness metric and in the case of a bad correction reverts back to a regular PDC-MAP correction. For this purpose, a regular MAP correction is performed for each target in addition to the msMAP correction. Then the goodness metric is calculated for both corrections and PDC decides for each target individually whether it should use the msMAP correction or revert to regular MAP. The process is illustrated in Figure 8.50. Decisions for each target are made on a quarterly basis, one quarter being one unit of work for PDC. Thus, a multi-quarter stitched light curve can have PDC data for individual quarters using both regular MAP and msMAP processed data. The threshold values can be set as input parameters to PDC. The vetting is biased toward preserving transit signals if a choice must be made (*Kepler* is primarily a transit-finding mission); however, for individual targets either regular MAP or msMAP may be more desirable for specific types of analysis.

By manual validation of the vetting results, we found that almost all of the bad corrections – about 90% – are successfully detected (using the threshold pa-



**Figure 8.50** The logic flow of the vetting process, which decides whether the regular MAP or msMAP fit is used. Includes a selection bias of 0.1 towards msMAP. From Figure 12 of Stumpe et al. (2014).

**Table 8.1** The confusion matrix of the vetting process for the 2919 targets of channel 7.3 in Q10.<sup>a</sup>

	msMAP correction <i>is</i> good	msMAP correction <i>is</i> bad	
correction <i>identified</i> as good	2775 (95.1%)	6 (0.2%)	2781 (95.3%)
correction <i>identified</i> as bad	95 (3.2%)	43 (1.5%)	138 (4.7%)
	2870 (98.3%)	49 (1.7%)	

<sup>a</sup>From Table 1 in Stumpe et al. (2014).

rameters shown in Figure 8.50), in which case PDC reverts back to the regular PDC-MAP correction. The confusion matrix of this vetting process is shown in Table 8.1 for the 2919 targets of channel 7.3 in Q10. Spot tests of other channels and quarters gave similar numbers. With this process, the number of bad corrections is reduced to only a handful ( $\sim 0.2\%$ ) per channel. For the false negatives that revert to regular MAP when the msMAP fit is actually good, we find the regular MAP fit is generally good as well and no harm is done reverting to regular MAP. It is interesting that for a large fraction of targets where regular MAP performs better than msMAP, the targets are highly variable. For such targets, msMAP has a tendency to either attenuate the stellar signals or introduce some high-frequency noise. These same targets are precisely the ones the original MAP method was designed to correct.

**8.6.4.2 Edge Effect Mitigation** An essential step in computing the wavelet transform of a discrete time series is circular convolution with a scaling filter (Percival & Walden, 2000). Wrapping the signal around from the end to the beginning gives rise to undesirable ‘edge effects’ at the boundaries, when there is discontinuity between the first and last parts of the time series. One way to mitigate these edge effects is to extend the time series by reflection, or to append a time-reversed copy of the time series at its right end (Percival & Walden, 2000). This ensures continuity, although it does tend to form cusps at the boundary, which

also give rise to edge effects. Another extension method is zero-padding, or adding zeros at the end of the signal to pad its length (Jensen & la Cour-Harbo, 2000). Zero-padding introduces edge effects due to the discontinuity between the ends of the signal and the zeros. When extension methods are used, the time series of wavelet coefficients and of the multiresolution analysis at each scale are subsequently truncated at the boundaries of the original time series.

Another concern regarding edge effects is their temporal extent. When a discrete time series is transformed to the wavelet domain, wavelet coefficients near the boundaries are unreliable, since they are influenced by extrapolated data (if an extension method is used) or by wrap-around (if there is no extension). There is a “zone of influence” in the timeseries of wavelet coefficients at each scale near the boundaries, the temporal extent of which increases with scale and also with the length of the wavelet filter (Percival & Walden, 2000). Wavelet coefficients at the largest scales are the most affected. The size of the zone of influence at all scales is reduced by using a shorter wavelet filter. We are currently using a Daubechies wavelet with a filter of length 12, which seems to adequately reduce the zone of influence.

We experimented with several extension methods to gauge their effectiveness in edge effect mitigation. We sought a method that would smoothly stitch together the signal and its extensions without the discontinuities and cusps that can give rise to severe edge effects. The method we settled on, which achieves this goal in most cases, involves extending the flux time series at each boundary using a sign-inverted, time-reversed copy of itself, explained in detail below.

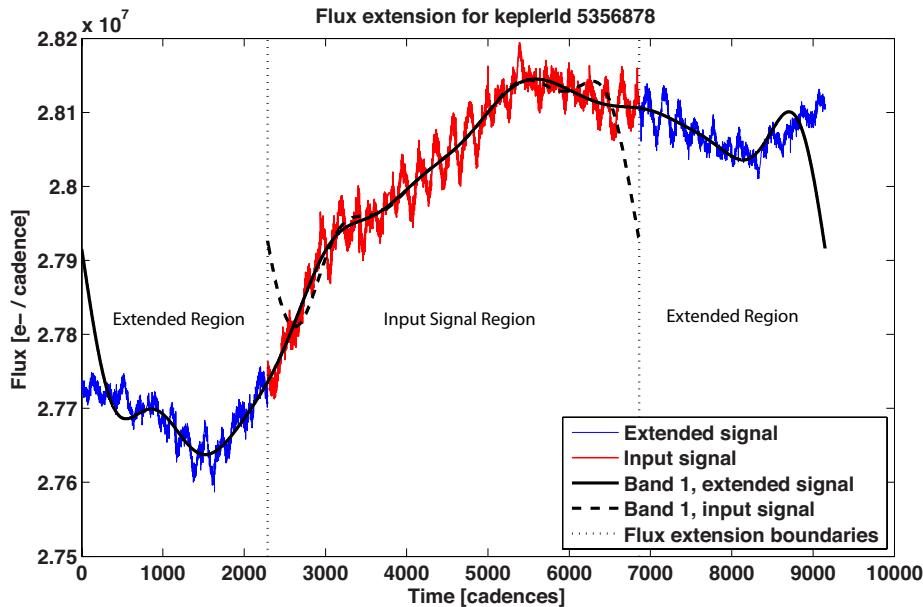
Matching up the extrapolated signal at the boundaries when the original signal contained an appreciable amount of noise or high-frequency variability introduced an added complication. We addressed the problem by estimating the flux-level at each boundary by low-pass filtering the 500 nearest cadences via linear interpolation. The right and left flux extensions are shifted vertically by the appropriate offsets so that they will match the estimated input flux at the boundaries. This makes the flux extension relatively robust to high-frequency variation or noise in the signal.

The extensions at the left and right boundaries are

$$L(t+T) = \left[ -G(t) + F(t_1) + \hat{F}(t_1) \right] r(t) + [F(t)] [1 - r(t)], \quad \text{and} \quad (8.46)$$

$$R(t-T) = \left[ -G(t) + F(t_2) + \hat{F}(t_2) \right] [1 - r(t)] + [F(t)] r(t), \quad (8.47)$$

where  $F(t)$  is the flux time series,  $t_1$  and  $t_2$  are its left and right boundaries,  $T$  is the length of the flux time series,  $G(t)$  is the reflected version of  $F$ ,  $\hat{F}(t_1)$  is the interpolated estimate of  $F(t_1)$ , and  $r(t) = (t - t_1) / (t_2 - t_1)$  is a linear ramp from 0 to 1 over the domain of the flux time series. The value of  $L$  at its right boundary (where it is to be stitched to the left boundary of the input flux) is  $L_{t_1} = \hat{F}(x_1)$ . Similarly, for  $R$ ,  $\hat{F}(t_2)$  is the interpolated estimate of  $F(t_2)$  and the value of  $R$  at its left boundary (where it is to be stitched to the right boundary of the input flux) is  $R_{t_2} = \hat{F}(x_2)$ . See Figure 8.51 for an example of the flux extension and subsequent band 1 light curve. Notice that the edges of the band 1 curve are well behaved.



**Figure 8.51** Edge effect mitigation showing the original light curve, the Band 1 light curve without extended regions, and then the edge effect mitigation extensions and the subsequent Band 1 light curve when utilizing the edge extensions. The Band 1 light curve with edge extensions is clearly well behaved at the light curve edges. From Figure 13 of Stumpe et al. (2014).

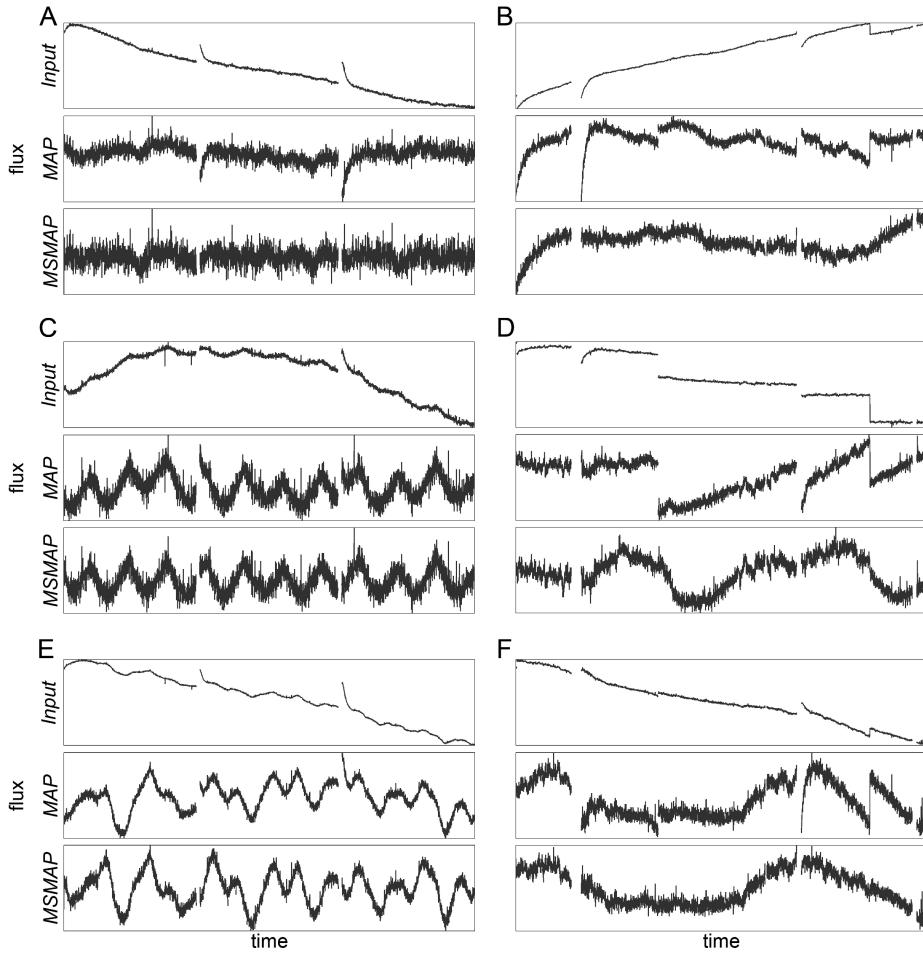
**8.6.4.3 Application to Short Cadence Data** The primary mission for *Kepler* is detection of Earth-like planets, and so development effort on the Pipeline emphasizes LC data. However, systematic error-corrected SC data from PDC is also provided to the users. The principal issue with applying the MAP technique to SC data is the limited number of targets per module output. No more than 512 SC targets are collected at any time and these are spread over the entire FOV so that the number of SC targets per channel is small and at most about a dozen. This is too small of a sample for the prior PDF or basis vectors to be properly formulated. However, all SC targets are also LC targets, and so priors are already developed for all SC targets. A simple way to extend MAP to SC data is to use the basis vectors interpolated from LC and also the long cadence fit coefficients as the prior as discussed in Subsection 8.5.7. However, multiscale MAP is not utilized for SC data.

## 8.6.5 Performance Evaluation

With the presented multiscale extension to PDC-MAP, we observe a significant improvement in correction quality compared to the previous version of PDC-MAP. In particular, the two main deficiencies of PDC-MAP, residual systematic errors and noise injection, are improved in msMAP.

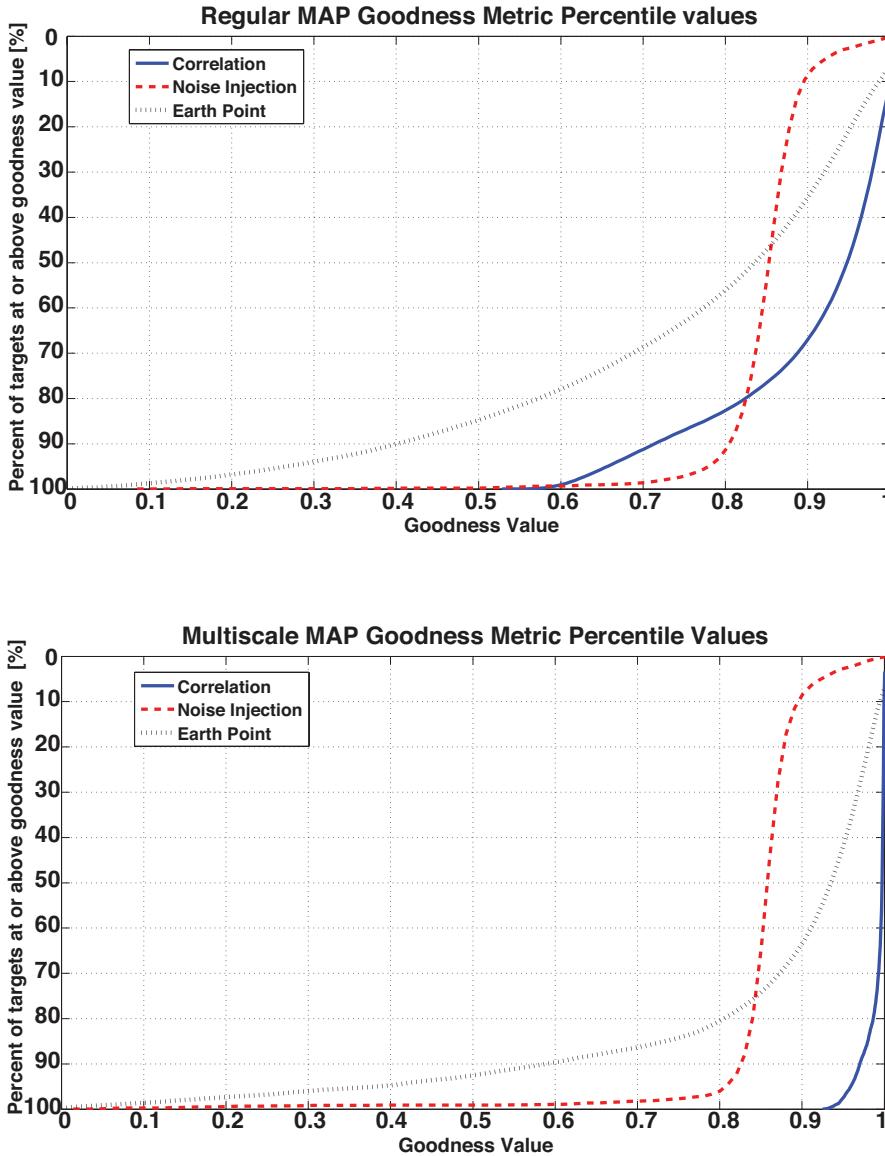
Figure 8.52 shows several example cases of light curves where regular MAP did not perform an optimal correction, but a substantially better correction with the new msMAP. Panels A and B show the initially discussed example from Figure 8.39. Note that we have deliberately picked cases where the regular PDC-MAP did not perform as well, which as a reminder, is only about every fifth light curve. In the other cases, the corrected time series of regular MAP and msMAP are usually very similar or almost identical.

The PDC goodness metric is invaluable in comparing cotrending methods. Here we can use it to compare the performance of a regular MAP run and a new multi-scale MAP run. Figure 8.53 shows the performance of regular MAP versus multi-scale MAP for the three goodness compo-



**Figure 8.52** Examples of msMAP correction performance improvements. The top row in each panel shows the PDC input time series, the middle row the PDC-MAP corrected time series, and the bottom row the msMAP corrected time series. Vertical scales are rescaled between panels to show detail. Panels A, C, E show examples with regular data quality (Q10 data), and panels B, D, F show cases with discontinuities due to attitude tweaks (Q2 data). The example in panel A and B are the same ones as in Figure 8.39. From Figure 1 of Stumpe et al. (2014). From Figure 14 of Stumpe et al. (2014).

nents: 1) Residual Correlation, 2) Injected Noise, and 3. Earth-point Recovery Removal. These figures are for Q10 module output 2.1 data. Goodness values are plotted as a “cumulative distribution function” where the horizontal axis gives the goodness value and the vertical axis gives the percent of targets with this goodness or above. The goodness metric is calibrated such that 0.8 or above is considered a “good” correction. There is a clear performance gain in the residual correlation and Earth-point components. There is a modest overall performance gain for the noise injection, and the “tail” of targets with noise goodness below 0.8 is reduced by over half. The stellar preservation goodness component does not change substantially and is not shown in the figure. The fraction of targets with residual correlation (defined by a correlation goodness below 0.8) has been reduced from about 20% down to near zero. The fraction of targets with residual Earth-point recoveries has been reduced from about 40% to 20%.



**Figure 8.53** Comparison of three of the goodness metric components between regular MAP and multi-scale MAP. The top plot is regular MAP the bottom is multi-scale MAP. From Figure 15 of Stumpe et al. (2014).

### 8.6.6 Outlook

One remaining issue is that the Earth-point recoveries have a very strong signal on the order of 150 cadences, which is the same scale as the oscillations from the three-day reaction wheel desaturation cycle (see Figure 8.2B and Figure 8.2C for illustrations of these errors). Consequently, these two systematic errors cannot be separated based on their characteristic scale. This is generally not a serious problem, but it does lead to one of those two errors not being perfectly corrected in some cases. One option to improve this situation is to explicitly add either of these signals as an additional basis vector. On first sight, the blatant artifact from the Earth-point

recoveries might seem like a good candidate to model with a simple exponential following the monthly Earth-point gap. However, their magnitude and shape can vary significantly between light curves, and therefore injecting an artificial basis vector for the whole set of light curves of one CCD channel would likely be non-ideal. Another choice would be to add a basis vector for the rather subtle 3-day reaction wheel desaturation cycle, but the resulting trend is not quite periodic or highly regular since its strength varies from the beginning to end of each quarter. One of the core principles of PDC-MAP has so far been to not use manually designed signals for the basis vectors but rather to generate them solely from the light curve data. However, augmentation to provide explicit known trends might prove useful and could be investigated in future work: an example is the explicit attitude tweak correction in Subsubsection III. There is also the option to perform the entire MAP fit in the transformed wavelet domain, whereas we currently transform back into the time domain after band-splitting. This may allow us to cleanly separate the reaction wheel desaturation from the Earth-point recoveries.

We would also desire to get a proper MAP fit working for the longest bands. There is no fundamental reason we cannot or should not. For the longest band the principal obstacles are resolving the issue with wavelet artifacts (see Subsubsection 8.6.3.3) and obtaining better long-period priors.

Perhaps the most promising candidate for further improvement is the generation of better priors for the MAP-fit. As has been discussed for the original version of PDC-MAP and again in this work (see Subsubsection 8.6.4.1), bad priors can sometimes lead to a very low correction quality. This problem is mitigated here in the new msMAP, because cases of poor performance are identified automatically by the goodness metric quality control and the better of the two corrections (msMAP or regular MAP) is used. However, even better would be to avoid these corrections in the first place. Furthermore, better priors would most likely also solve the problem where using a MAP-fit in the longest-scale band can lead to an artifact.

## 8.7 Conclusions

We have presented the *Kepler* Presearch Data Conditioning module, a central part of the *Kepler* Science Processing Pipeline that is tasked with correcting systematic errors in the light curves. Our approach, using wavelet-based bandsplitting to decouple errors on different scales and perform a multiscale correction, generates high-quality light curves. In a sense, the correction characteristics of PDC can be regarded as the best of both worlds. The first version of PDC-LS (Twicken et al., 2010a) only rarely exhibited residual systematic errors in the light curves but was prone to overfitting and removing astrophysical features. The single-scale PDC-MAP represented a milestone improvement and was not sensitive to overfitting at all, but was sometimes too conservative in preserving residual systematics in the light curves. The performance of PDC can be tuned to both of these extremes by setting the parameters accordingly, in particular the number of bands. Between these extremes, we found that there is an optimal parameter set (i.e. 3 or 4 bands) that delivers excellent correction of systematic errors without removal of astrophysical signals and without injection of high-frequency noise. Further, our extensive testing of the parameter range showed that this optimal parameter set is sufficiently broad and robust and that excellent correction quality can be achieved for almost all CCD channels and data quarters without the need for individual fine-tuning of the parameters for each situation. Rare cases where the multiscale correction fail are caught by an automatic quality control process via the PDC goodness metric, and the single-scale MAP is used instead. In summary, PDC is capable of producing error-corrected light curves of unprecedented quality from which we expect users of the *Kepler* data to benefit, furthering the study of stellar astrophysics as well as the search for extrasolar planets.

## Bibliography

- Beck, P. G., Bedding, T. R., Mosser, B., et al., 2011. “Kepler Detected Gravity-Mode Period Spacings in a Red Giant Star,” *Science*, 332, 205
- Borucki, W. J., Koch, D., Basri, G., et al., 2010. “Kepler Planet-Detection Mission: Introduction and First Results,” *Science*, 327, 977
- Borucki, W. J., Koch, D. G., Batalha, N., et al., 2012. “Kepler-22b: A 2.4 Earth-Radius Planet in the Habitable Zone of a Sun-like Star,” *ApJ*, 745, 120
- Bowman, A. W., & Azzalini, A. 1987. *Applied Smoothing Techniques for Data Analysis* (Oxford University Press)
- Chaplin, W. J., Kjeldsen, H., Christensen-Dalsgaard, J., et al., 2011. “Ensemble Asteroseismology of Solar-Type Stars with the NASA Kepler Mission,” *Science*, 332, 213
- Clarke, B. D., Allen, C., Bryson, S. T., et al. 2010. “A Framework for Propagation of Uncertainties in the Kepler Data Analysis Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774020
- Cohen, A., Daubechies, I., & Feauveau, J.-C., 1992. “Biorthogonal Bases of Compactly Supported Wavelets,” *Communications on Pure and Applied Mathematics*, 45, 485
- D’Agostini, G. 2003. *Bayesian Reasoning in Data Analysis* (World Scientific)
- DeGroot, M. 1970. *Optimal Statistical Decisions* (Wiley-Interscience)
- Donoho, D. L., & Johnstone, I. M., 1994. “Ideal Denoising in an Orthonormal Basis Chosen from a Library of Bases, journal = Comptes Rendus Acad. Sci., Ser. I,”, 319, 1317
- Doyle, L. R., Carter, J. A., Fabrycky, D. C., et al., 2011. “Kepler-16: A Transiting Circumbinary Planet,” *Science*, 333, 1602
- Fressin, F., Torres, G., Rowe, J. F., et al., 2012. “Two Earth-sized planets orbiting Kepler-20,” *Nature*, 482, 195
- Haas, M. R., Batalha, N. M., Bryson, S. T., et al., 2010. “Kepler Science Operations,” *ApJL*, 713, L115
- Holman, M. J., Fabrycky, D. C., Ragozzine, D., et al., 2010. “Kepler-9: A System of Multiple Planets Transiting a Sun-Like Star, Confirmed by Timing Variations,” *Science*, 330, 51
- Jenkins, J. M., 2002. “The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets,” *ApJ*, 575, 493
- Jenkins, J. M., Caldwell, D. A., & Borucki, W. J., 2002. “Some Tests to Establish Confidence in Planets Discovered by Transit Photometry,” *ApJ*, 564, 495
- Jenkins, J. M., Smith, J. C., Tenenbaum, P., Twicken, J. D., & Van Cleve, J. 2012. “Planet Detection: The Kepler Mission,” in *Advances in Machine Learning and Data Mining for Astronomy*, ed. M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (Chapman and Hall, CRC Press), 355–381
- Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al., 2010. “Initial Characteristics of Kepler Long Cadence Data for Detecting Transiting Planets,” *ApJL*, 713, L120

- , 2010. “Overview of the Kepler Science Processing Pipeline,” *ApJL*, 713, L87
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010c. “Transiting Planet Search in the Kepler Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D
- Jensen, A., & la Cour-Harbo, A. 2000. *Ripples in Mathematics: The Discrete Wavelet Transform* (Springer)
- Kay, S. M. 2012. *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice Hall)
- Koch, D. G., Borucki, W. J., Basri, G., et al., 2010. “Kepler Mission Design, Realized Photometric Performance, and Early Science,” *ApJL*, 713, L79
- Kolodziejczak, J., & Morris, R. L. 2012. “Methods for Detection and Correction of Sudden Pixel Sensitivity Drops,” KADN 26304, NASA Ames Research Center Kepler Mission
- Kovács, G., Bakos, G., & Noyes, R. W., 2005. “A Trend Filtering Algorithm for Wide-Field Variability Surveys,” *MNRAS*, 356, 557
- Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al., 2011. “A Closely Packed System of Low-Mass, Low-Density Planets Transiting Kepler-11,” *Nature*, 470, 53
- Meibom, S., Barnes, S. A., Latham, D. W., et al., 2011. “The Kepler Cluster Study: Stellar Rotation in NGC 6811,” *ApJL*, 733, L9
- Minka, T. P. 2008. “Automatic Choice of Dimensionality for PCA,” Tech. Rep. 514, MIT Media Lab. Perceptual Computing Section Technical Report
- Percival, D. B., & Walden, A. T. 2000. *Wavelet Methods for Time Series Analysis* (Cambridge University Press)
- Savitzky, A., & Golay, M. J. E., 1964. “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *Analytical Chemistry*, 36, 1627
- Smith, J. C., Stumpe, M. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning II – A Bayesian Approach to Systematic Error Correction,” *PASP*, 124, 1000
- Stello, D., Meibom, S., Gilliland, R. L., et al., 2011. “An Asteroseismic Membership Study of the Red Giants in Three Open Clusters Observed by Kepler: NGC 6791, NGC 6819, and NGC 6811,” *ApJ*, 739, 13
- Stumpe, M. C., Smith, J. C., Catanzarite, J. H., et al., 2014. “Multiscale Systematic Error Correction via Wavelet-Based Bandsplitting in Kepler Data,” *PASP*, 126, 100
- Stumpe, M. C., Smith, J. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning I – Architecture and Algorithms for Error Correction in Kepler Light Curves,” *PASP*, 124, 985
- Tamuz, O., Mazeh, T., & Zucker, S., 2005. “Correcting Systematic Effects in a Large Set of Photometric Light Curves,” *MNRAS*, 356, 1466
- Tenenbaum, P., Bryson, S. T., Chandrasekaran, H., et al. 2010. “An Algorithm for the Fitting of Planet Models to Kepler Light Curves,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400J

- Twicken, J. D., Chandrasekaran, H., Jenkins, J. M., et al. 2010a. "Presearch Data Conditioning in the Kepler Science Operations Center Pipeline," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401U
- Twicken, J. D., Clarke, B. D., Bryson, S. T., et al. 2010b. "Photometric Analysis in the Kepler Science Operations Center Pipeline," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 774023
- Van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Instrument Handbook: (KSCI-29033-002) (Moffett Field, CA: NASA Ames Research Center)
- Vetterli, M., & Kovacevic, J. 1995. Wavelets and Subband Coding (Prentice Hall)
- Welsh, W. F., Orosz, J. A., Carter, J. A., et al., 2012. "Transiting Circumbinary Planets Kepler-34 b and Kepler-35 b," Nature, 481, 475
- Witteborn, F. C., Van Cleve, J., Borucki, W., Argabright, V., & Hascall, P. 2011. "DEBRIS Sightings in the Kepler Field," in Proc. SPIE, Vol. 8151, Techniques and Instrumentation for Detection of Exoplanets V, 815117
- Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S., 2009. "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, journal = IEEE Trans. Pattern Anal. Mach. Intell.", 31, 39



## **PART III**

---

# **THE KEPLER TRANSIT SEARCH ENGINE**



# CHAPTER 9

---

## TRANSITING PLANET SEARCH

JON M. JENKINS<sup>1</sup>, PETER TENENBAUM<sup>2</sup>, SHAWN SEADER<sup>2</sup>, CHRISTOPHER J. BURKE<sup>2</sup>, SEAN D. McCUAULIFF<sup>3</sup>, JEFFREY C. SMITH<sup>2</sup>, JOSEPH D. TWICKEN<sup>2</sup>, AND HEMA CHANDRASEKARAN<sup>2</sup>

<sup>1</sup>NASA Ames Research Center, Moffett Field, CA 94035, <sup>2</sup>The SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035, <sup>3</sup>Wyle Labs/NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** The *Kepler* Mission simultaneously measured the brightnesses of  $\sim$ 165,000 stars every 29.4 minutes during each of seventeen 93-day “quarters” over a 4-year mission with the aim to discover Earth-size planets transiting Sun-like stars in the habitable zone.<sup>1</sup> The Transiting Planet Search (TPS) component of the SOC Science Pipeline conducts the actual search for signatures of transiting planets in the systematic error-corrected light curves. Potential transiting planets identified by TPS are subjected to further analysis and scrutiny by the Data Validation (DV) pipeline component. Detecting transits is a signal detection problem in which the signal of interest is a periodic pulse train and the predominant noise source is a non-white, non-stationary 1/f-type process of stellar variability. The situation is complicated by the fact that many stars exhibit coherent or quasi-coherent oscillations. In addition, instrumental effects pose a substantial source of false alarms, which motivated substantive effort to develop mitigations in TPS to screen out anomalous detections while preserving a high detection rate for true planetary signatures. This paper details the results of that effort, which culminated in the SOC 9.3 version of TPS and delivered 34,032 potential transiting planet signatures to the NExSci Exoplanet Archive.

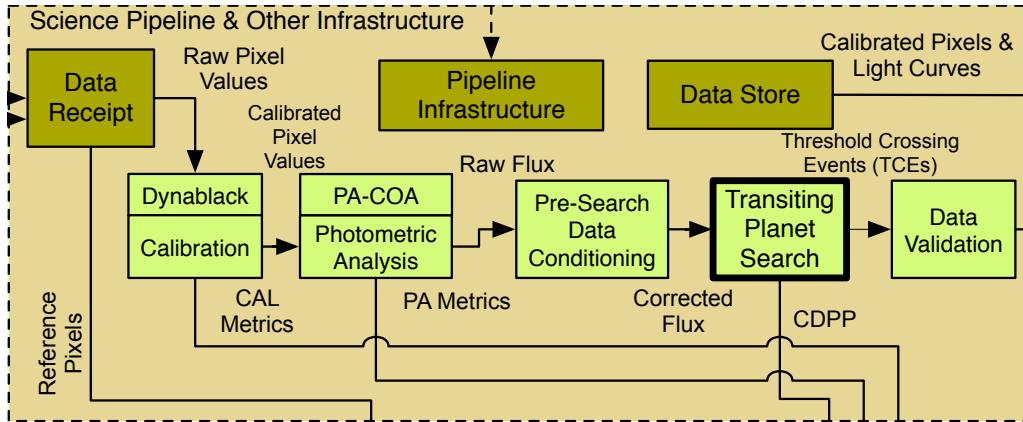
**Keywords:** *Kepler* Mission, exoplanet, transit, detection algorithm

### 9.1 Introduction

The Transiting Planet Search (TPS) module is responsible for identifying transit-like features in the systematic error-corrected, long cadence (LC) flux time series produced by the Photometry Pipeline (see Part II). Detecting  $\sim$ 100 ppm-deep transit signatures of Earth-size planets is a daunting task even for the exquisite *Kepler* photometric dataset, where the typical observation noise is  $\sim$ 30 ppm on 6.5-hour timescales for a  $K_p = 12$  main-sequence star (Gilliland et al., 2011, 2015). The transit detection algorithm must contend with outliers, highly variable stars such as flare stars and giant stars, radiation-induced transients, and instrumental effects such as rolling bands (see Chapter 4) that can mimic transit signatures from the standpoint of a linear

<sup>1</sup>The first quarter, Q1, was only 34 days long due to the launch date and commissioning period. The last quarter, Q17, was only 31 days long due to the mission-ending loss of reaction wheel #4 and contained a 10-day rest period to attempt to increase the lifetime of this reaction wheel.

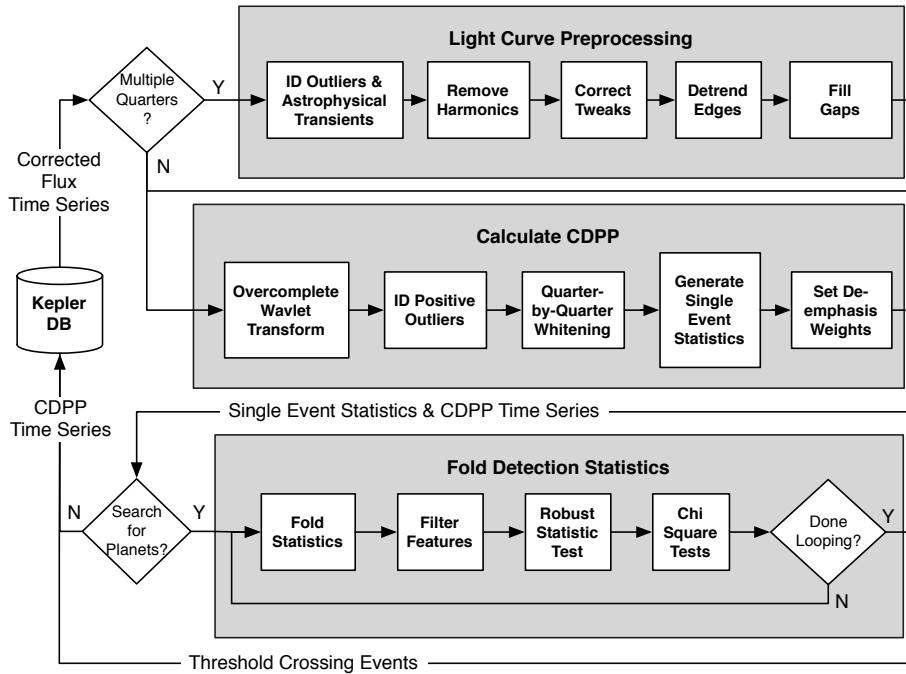
transit detection algorithm. Figure 9.1 shows TPS in the context of the SOC Science Processing Pipeline.



**Figure 9.1** Data flow diagram for the SOC Science Processing Pipeline. TPS in the context of the architecture of the SOC. TPS searches the systematic error-corrected light curves produced by the Presearch Data Conditioning (PDC) module for transit-like features that are then subjected to physical modeling and a suite of diagnostic tests in the Data Validation (DV) module. DV also calls TPS internally to enable a search for multiple planets within each light curve.

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, Dynablack (DYN – see Chapter 4) and Calibration (CAL – see Chapter 5) calibrate 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, Photometric Analysis (PA – see Chapter 6 and Chapter 7) 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 containing each target star image. Third, Presearch Data Conditioning (PDC – see Chapter 8) identifies and removes signatures of systematic effects in the photometric time series, such as changes in pointing or focus, and fills any intraquarter gaps to condition the time series for TPS. TPS fills the inter-quarter gaps then searches the corrected flux time series for signatures of periodic pulse trains indicative of transiting planets. Potential transit signals are designated threshold crossing events (TCEs) and are subsequently examined in detail by the Data Validation (DV – see Chapter 11 and Chapter 12) component to establish or break confidence in these transit-like features as planetary signatures. TPS is also called within DV to iteratively search for additional planetary signatures in a light curve after each TCE is analyzed and its transit pulses are removed from the light curve.

As illustrated in Figure 9.2, 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 Jenkins et al. (2002). 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 half the length of the time series.



**Figure 9.2** Block diagram for TPS, which has three major blocks of functionality. The first major block of TPS preprocesses the light curves to identify residual outliers and astrophysical transients, identify and remove harmonic signatures, correct for pointing tweak-induced discontinuities, detrend edge effects, and finally, fill gaps created by these steps as well as intra-quarter gaps. The second major block of TPS transforms the light curve into the wavelet domain to analyze the time-varying characteristics of the observation noise and to generate the single event statistics, which indicate the likelihood that a transit of the given pulse duration is occurring at any given time in the light curve. A by-product of the generation of the single-event statistics is Combined Differential Photometric Precision (CDPP), a measure of the photometric precision for each star on transit timescales. When run in planet search mode, the single event statistics are folded over trial orbital periods to test the significance of repeating transit-like features. Stars for which these multiple-event statistics (MES) exceed  $7.1\sigma$  and that pass a set of statistical tests are designated Threshold Crossing Events (TCEs) and are persisted to the *Kepler* File Store, along with the CDPP time series and other information, such as the epoch and period of the most likely transit pulse train. For monthly datasets, TPS measures the photometric precision achieved for as many as 169,000 target stars, in which case the third subcomponent was skipped. This flow is executed independently for each trial transit pulse duration, which varied from 1 to 15 hours.

TPS applies an adaptive, wavelet-based matched filter to detect potential transiting signatures (Jenkins, 2002). The detector analyzes the power spectral density (PSD) of the observation noise as a function of time to design and implement a whitening filter. The detector then applies the whitening filter, flattening the PSD and transforming the observation noise into white Gaussian noise to a good approximation. TPS applies a simple matched filter against the whitened transits to formulate single event statistics representing the likelihood that a transit of a given duration is occurring at each time step. These single event statistics are folded over trial orbital periods to formulate the multiple event statistics as a function of period and phase. This process is repeated for 14 separate pulse durations ranging from 1.5 to 15 hours long.

When the detection statistic for a trial transit signature exceeds the  $7.1\sigma$  threshold,<sup>2</sup> it is subjected to additional statistical tests within TPS that are designed to eliminate spurious false alarms while maximizing the sensitivity of the transit search to true planetary signals (Seader et al., 2013). These tests include a robust version of the multiple-event statistic (*MES*) that keyed off the detection in the first place, and two chi-square vetoes, dubbed  $\chi^2_{(2)}$  and  $\chi^2_{GOF}$  for "Goodness of Fit" (Seader et al., 2013). If the event is discarded, TPS "notches" out the period and epoch associated with the failed TCE and examines up to 1,000 other multiple event statistics that exceed  $7.1\sigma$  in descending order. If a detection is caused by a single strong feature that does not fold onto comparably strong features, it is masked out of the time series. Up to two such features can be identified and removed from the light curve. Signatures that pass all three of these additional tests are recorded as TCEs and then submitted to DV for physical modeling and additional diagnostic tests.

Monitoring the photometric precision obtained by *Kepler* has also been a high priority and was obtained on a monthly (and quarterly) 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 inverse signal-to-noise ratio (SNR) of a 1 ppm-deep transit of a given duration (Jenkins et al., 2010b; Christiansen et al., 2012). This is essentially the effective photometric noise "seen" by a transit of a given duration from a detection point of view. That is, a 20 ppm CDPP at 6.5 hours indicates that a 100 ppm-deep 6.5-hour transit would be expected to produce a  $5\sigma$  detection. Each month, CDPP, along with a suite of performance metrics developed during processing as the data proceed through the pipeline, was monitored and reported by the Photometric Performance Assessment (PPA) component (Li et al., 2010). The SOC was required to monitor CDPP for transit durations of 3, 6, and 12 hours as key indicators of *Kepler*'s sensitivity to small, rocky transiting planets. The typical duration of transit varies from a few hours for close-in planets to 16 hours for a Mars-size orbit (Koch et al., 2010). 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.

A number of issues identified since science operations commenced on May 12, 2009 have required significant modifications to the Science Pipeline and to TPS. Indeed, the development of TPS continued throughout the *Kepler* flight operations and into the close-out of the mission. Many of these improvements have been periodically reported in Jenkins et al. (2010b) and the articles documenting the TCE catalogs (Tenenbaum et al., 2012, 2013, 2014; Seader et al., 2015; Twicken et al., 2016), but there have been several major enhancements to TPS in the final software development cycle for the SOC build 9.3. Some of the most important modifications include a modification to the method by which the noise power is estimated in the whitener, the use of quarter-by-quarter whitening to reduce edge effects and associated false positives, and a modification in the long gap fill algorithm to reduce edge effects and bias in the noise power estimates. This paper describes the final, as-built SOC 9.3 version of TPS and gives special attention to the

<sup>2</sup>The threshold of  $7.1\sigma$  was established by a Monte Carlo approach to establish the threshold required to control the false alarm rate due to statistical fluctuations to  $<1$  for the *Kepler* campaign, assuming the observation noise is well modeled as broadband, colored Gaussian noise (Jenkins et al., 2002).

latest developments incorporated therein. We also describe the performance of TPS for the final *Kepler* transiting planet search through all 17 quarters of data, called Data Release 25 (DR25).

This article is organized as follows: Light curve preprocessing is discussed in Section 9.2. Section 9.3 discusses the joint time-frequency domain characterization of the observation noise, the generation of the single event statistics and photometric precision metrics, and summarizes the core detection algorithm. Section 9.4 describes the folding of the detection statistics, the application of the vetoes, and the iterative transit search. Section 9.5 discusses the performance of TPS for the final *Kepler* transiting planet search. Concluding remarks are given in Section 9.6.

## 9.2 Light Curve Preprocessing

As effective as PDC is at identifying and removing instrumental signatures and outliers, it is necessary for TPS to make an additional pass at removing these effects from the light curves prior to the transit search in order to minimize the number of spurious detections due to edge effects and residual instrumental transients. TPS can be more aggressive at doing so than PDC, often at the expense of other, non-transit astrophysical signatures, since the light curves TPS searches are not archived to MAST for the astronomical community.

The first stage of TPS conditions each quarterly light curve to remove residual instrumental effects, and non-transit astrophysical signatures that tend to generate false alarms, normalizes the baseline in each quarter, fills both intra- and inter-quarter gaps, and stitches the quarterly light curves together into one single flux time series. This stage also identifies eclipsing binary and giant planetary transit signatures to protect them from being removed by these filters.

### 9.2.1 Identification of Outliers and Astrophysical Transients

TPS identifies signatures of eclipsing binaries and deep planetary transits by robustly fitting a piecewise polynomial to each quarterly light curve and looking for points that are more than 10 median absolute deviations (MAD) below the trend line. The order of the polynomial fitted to each 72-hour interval is set by Akaike's Information Criterion (AIC—Akaike, 1974). If the points identified are isolated, they are identified as outliers and removed from the light curve. If they are not isolated outliers, the locations are saved and used to selectively ignore these events in the next step, where harmonic content is identified and removed, and in the formulation of the CDPP estimates.

In addition to identifying eclipses and strong transit features, TPS inverts the light curves and performs the same analysis to identify potential microlensing events, which are similarly protected in the next processing step.

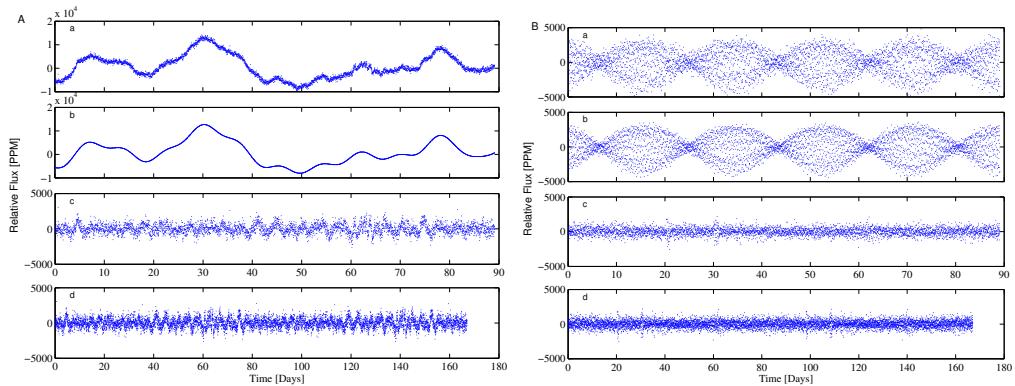
### 9.2.2 Identifying and Removing Phase-Shifting Harmonics

The adaptive, wavelet-based matched filter employed by TPS to search for transiting planet signals is not well suited to compact signals in the frequency domain, such as sinusoidal signals due to highly periodic pulsations. Therefore, harmonic signals are fitted and removed from each light curve on a quarter-by-quarter basis prior to conducting the transit search. This process significantly reduces the number of false alarms that would result from retaining these harmonic signatures but may also degrade or remove short-period ( $< 3$  days) transit signals (Christiansen et al., 2013, 2015). The harmonic fitting is conducted iteratively and is numerically intensive; a maximum number of harmonic components are therefore permitted to be fitted and removed in order to manage the time spent on this process and also to avoid overfitting.

Once the deep transits and eclipses have been identified (as described in Subsection 9.2.1), the cadences containing such events are temporarily filled using an autocorrelation-based short

data gap fill algorithm (Chandrasekaran, 2004). The time series is extended to the next power of 2 (in length) using the approach of Jenkins et al. (2002) and a Hanning window-weighted periodogram is formed.<sup>3</sup> 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 (47-point) median filter is applied to the periodogram and then the result is smoothed with a (47-point) moving average window.<sup>4</sup> The median filter ignores isolated peaks in the periodogram. Next, this background PSD estimate is divided point-wise 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 be computationally intensive.

Figure 9.3 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.



**Figure 9.3** Harmonic removal and extension of two flux time series. A: A star with low-frequency oscillations. B: A star with high-frequency oscillations and amplitude modulation. a (left and right): Original flux time series. b (left and right): Detected harmonic signature. c (left and right): Flux time series with harmonics subtracted. d (left and right): Harmonic-free time series extended to 8192 samples. From Figure 3 of Jenkins et al. (2010b).

Previous versions of the software allowed a fixed maximum number of harmonics (25) to be removed from each quarterly light curve, regardless of the length of the quarter. This led to inconsistent fitting of periodic stellar variability wherein short quarters (e.g., Q1, Q4 for stars on CCD Module 3, and Q17) had significantly more harmonic content removed, artificially reducing the apparent observation noise and thereby biasing the transit search.<sup>5</sup> The SOC 9.3 codebase adjusts the maximum number of harmonic components fitted in each quarter to be proportional to the length of the quarterly dataset, leading to more consistent harmonic fitting results across the full four-year dataset. The maximum number of harmonic components removed per target and quarter in the DR25 run ranged from 8 in Q17 and 9 in Q1 at the low end, to 25 in Q9, Q11, Q14, and Q15.

<sup>3</sup>The times series must be extended to a length equal to a power of 2 to allow for the use of FFTs. It is also essential to conserve the power spectral distribution while doing so, thus motivating the use of an autocorrelation-based approach.

<sup>4</sup>The window lengths for median and moving average filters are programmable and had values of 47 for the DR 25 run.

<sup>5</sup>Q1 began on May 13, 2009 immediately after the commissioning activities were completed and was only 33.46 days long. Module 3 died on January 9 2010, ~10 days into Q4. The second reaction wheel failed on May 11 2013 during the second month of Q17. In addition, Q17 has a 10 day gap imposed by an attempt to improve the reaction wheel's health with a rest period prior to its ultimate failure.

### 9.2.3 Edge Detrending of Contiguous Blocks of Flight Data

The *Kepler* photometer is sensitive to changes in its thermal state that can induce changes in the shape of the telescope and thereby introduce focus changes that affect the photometry. While the Earth-trailing, heliospheric orbit is extremely benign from a thermal standpoint, the spacecraft turned towards Earth every month to downlink data and performed a 90° rotation about its bore-sight every three months in order to reorient the sunshade and solar panels. When the spacecraft returned to the science attitude, it experienced thermal transients due to the attitude changes and the time spent at Earth-point (or in safe mode, in the case of unplanned departures from science attitude).

To remove trends at the ends of single-quarter data segments, TPS performs a robust fit of the form:

$$y = P_1 \exp(-x/P_2) + P_3 x + P_4 + P_5 \exp[(x-1)/P_6], \quad (9.1)$$

where  $y$  is the median-corrected flux,  $x$  is the sample time normalized to the unit interval [0, 1], and  $P_1$  through  $P_6$  are the parameters of the fit. This is then subtracted from the light curve. Equation 9.1 fits a line plus two exponential edge trends, one at the leading edge of the data region and one at the trailing region, with both the amplitude and the time constant of the exponentials as fit parameters. The form in Equation 9.1 was found to match the actual edge trends as well as the constrained polynomial fit which had been used in early versions of the software. The advantages of the reformulated edge-trend removal are: a reduced number of assumptions and/or configuration parameters for the fit, use of the full data segment for the entire fit, robust fitting, and the fact that this fit cannot introduce a polynomial “wave” into the data segment in an attempt to correct the edges (i.e., over fitting). Additionally, whereas in the past the edge detrending was applied only to full quarters of data, in the current implementation it is applied at any time when there was an interruption of data acquisition to change the spacecraft orientation. This was done to mitigate the thermal transients that occur when the spacecraft attitude is changed.

**9.2.3.1 Normalization** Because the PSF and the sensitivity of the CCD varies across the focal plane, the mean flux level for a target star can vary as a function of the observing season from quarter to quarter. To partially mitigate these effects, TPS normalizes each quarterly light curve segment by its median and then subtracts 1.

**9.2.3.2 Long Gap Filling** The quarterly segments are then stitched together prior to the planetary search. Because the Fast Fourier Transforms (FFT) is used by TPS to implement the Over-complete Wavelet Transform (OWT), it is necessary to fill all gaps to prevent the occurrence of artifacts in the joint time-frequency domain in which the observation noise is characterized and the transit detection is performed. Another important change to TPS in SOC 9.3 is a modification to the long gap fill algorithm. The new algorithm applies a narrow sigmoid taper for the periodic extension of time series across long data gaps ( $> 2.5$  days) prior to application of the FFTs for the wavelet filter bank. The sigmoid occupies the central 10% of the gap to be filled; data are simply reflected from the left and right hand sides of the gap, then weighted by the sigmoid taper and added together in the central 10% region. Note that any eclipses or deep planetary transits identified in the first step of this stage are not allowed to be reflected into the filled data gaps, which avoids false alarms due to “ringing” of the single event statistics from eclipses or deep transits in the gap filled data to the valid data outside the gaps. The former long gap fill algorithm used a linear taper across the entire gap, resulting in a systematic reduction in noise power estimates near gap edges. This artificial drop in noise power was tolerable when the entire four-year time series was whitened at once but resulted in significant biases in the sensitivity to transits or transit-like features near quarterly boundaries with the introduction of quarter-by-quarter whitening in SOC 9.3. Modifying the long gap fill algorithm also improved the performance of TPS for long data gaps within quarterly datasets.

After the long gap filling is complete, the quarter-stitched light curves are ready for the next processing stage.

### 9.3 Generation of Single Event Statistics and CDPP

Once the preprocessing is complete, TPS transforms the light curve into a joint time-frequency domain to analyze the PSD as a function of time to formulate a time-varying whitening filter and generate the single event statistics. This also provides an opportunity to identify residual positive impulsive outliers that have in the past generated significant numbers of spurious detections due to “ringing” effects in the wavelet domain. We begin this section with an overview of the wavelet-based matched filter.

#### 9.3.1 A Wavelet-Based Matched Filter

While the transit detection is not performed until the third stage of TPS, the generation of the CDPP time series and the formulation of the single event statistics can be understood only in the context of the detection algorithm. Thus we provide a full explication of the wavelet-based adaptive matched filter in this section.

The optimal detector for a deterministic signal in colored Gaussian noise is a pre-whitening filter followed by a simple matched filter (Van Trees, 1968; Kay, 1999). In TPS we implement a wavelet-based matched filter as per Jenkins et al. (2002) using Daubechies’ 12-tap wavelets (Daubechies, 1988). 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 observation 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 approximately 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  $x(n)$  be a flux time series. The OWT of  $x(n)$  is given by

$$\mathbb{W}\{x(n)\} = \{x_1(n), x_2(n), \dots, x_M(n)\}, \quad (9.2)$$

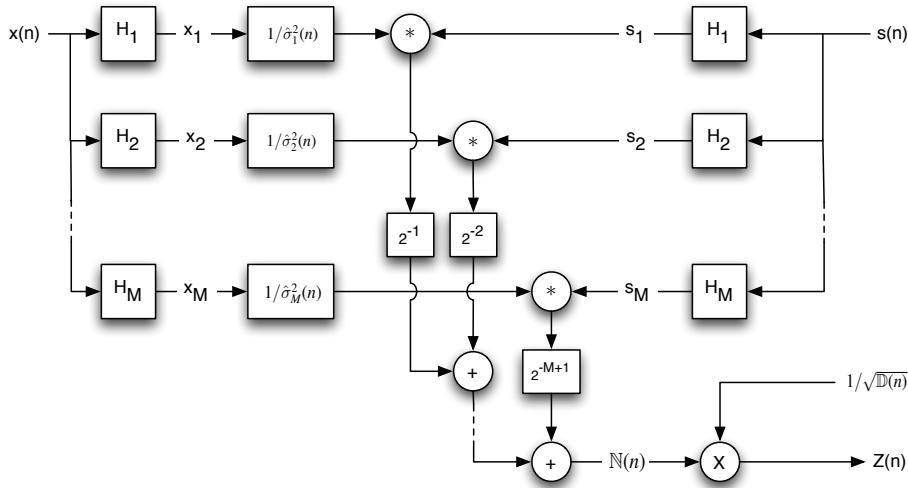
where

$$x_i(n) = h_i(n) * x(n), \quad i = 1, 2, \dots, M, \quad (9.3)$$

$*$  denotes convolution, and  $h_i(n)$  for  $i = 1, \dots, M$  are the impulse responses of the filters in the filter bank implementation of the wavelet expansion with corresponding frequency responses,  $H_i(\omega)$ , for  $i = 1, \dots, M$ . The filters are generated recursively as described in Appendix 9-A.

Figure 9.4 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 9.5. 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,  $\{x_i(n)\}$ ,  $i = 1, \dots, M$ , 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 with an analysis window chosen to be significantly longer than the duration of the trial transit pulse ( $7 \times$  the trial transit pulse duration for DR 25). Note that the analysis window for estimating the noise power increases by a factor of two from bandpass to bandpass, until the window covers the entire time series.

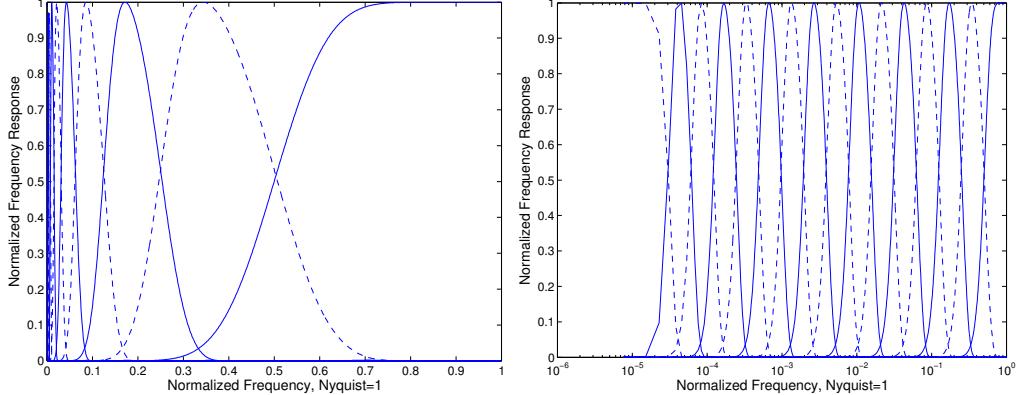


**Figure 9.4** Signal flow diagram for TPS. The wavelet-based matched filter is implemented as a filter bank with bandpass filters  $H_1, \dots, H_M$  progressing from high frequencies to low frequencies. The flux time series,  $x(n)$ , is expanded into  $M$  time series  $x_i(n)$ , for  $i = 1, \dots, M$ . Noise power,  $\hat{\sigma}_i^2(n)$ ,  $i = 1, \dots, M$  is estimated for each bandpass and then divided into the channel time series,  $x_i(n)$ , in order to “doubly” whiten  $x(n)$  in the wavelet domain. The trial transit pulse,  $s(n)$ , is processed through a copy of the filter bank and convolved with the “doubly” pre-whitened flux time series,  $x_i(n)/\hat{\sigma}_i^2(n)$ , in each bandpass. Since the whitening is a property of the observed noise process, normalizing the flux time series in each bandpass by the noise variance estimates effectively whitens both the signal of interest,  $s(n)$ , as well as  $x(n)$ , since the whitened version of  $x_i(n)$  and  $s_i(n)$  are  $x_i(n)/\hat{\sigma}_i(n)$  and  $s_i(n)/\hat{\sigma}_i(n)$ , respectively. Parseval’s theorem for undecimated wavelet representations allows us to combine the results for each bandpass together to form the numerator term,  $N(n)$  of Equation 9.6. A similar filterbank arrangement is used to furnish  $D(n)$  from Equation 9.6 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,  $Z(n)$ , is obtained by dividing the correlation term,  $N(n)$ , by the square root of the denominator term,  $D(n)$ . From Figure 4 of Jenkins et al. (2010b).

One of the most important improvements to TPS in SOC 9.3 is the incorporation of a non-decimated (in time) moving MAD filter for estimating the rms noise power time series,  $\hat{\sigma}_i(n)$ , in each of the wavelet filter bank’s band passes. This change was motivated by the observation that there was a measurable and significant duration dependent bias in the noise power estimates; the noise power for short duration transits was underestimated relative to that for long duration transits. Prior to SOC 9.3, TPS employed a decimated moving MAD filter due to computational run time constraints, so the moving MAD was not calculated for each sample in each of the wavelet band passes. We were able to implement the moving MAD filter algorithm more effi-

ciently and perform the computation without decimation; this eliminated the bias in the noise power estimates while maintaining adequate computational throughput.

These changes to TPS increase the sensitivity to transiting planet signatures, improve the uniformity of the sensitivity of the search, and enhance the performance and characteristics of the statistical bootstrap analysis (Jenkins et al., 2015) performed for each TCE in Data Validation.



**Figure 9.5** Frequency responses of the filters in the octave-band filterbank for a wavelet expansion corresponding to the signal flow graph in Figure 9.4 using Daubechies’ 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. From Figure 5 of Jenkins et al. (2010b).

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

$$Z = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{s}}}{\sqrt{\tilde{\mathbf{s}} \cdot \tilde{\mathbf{s}}}} = \frac{\sum_{i=1}^M 2^{-\min(i,M-1)} \sum_{n=1}^N [x_i(n)/\hat{\sigma}_i(n)] [s_i(n)/\hat{\sigma}_i(n)]}{\sqrt{\sum_{i=1}^M 2^{-\min(i,M-1)} \sum_{n=1}^N s_i^2(n)/\hat{\sigma}_i^2(n)}}, \quad (9.4)$$

where the time-varying channel variance estimates are given by taking the square of a moving MAD estimate with a telescoping window size:

$$\hat{\sigma}_i^2(n) = 1.4826 \text{ MAD} \{x(n - 2^{i-1}K), x(n - 2^{i-1}K + 1), \dots, x(n + 2^{i-1}K)\} \quad (9.5)$$

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 typically set to 7 times the trial transit duration. The factor of 1.4826 accounts for the bias between the MAD and the standard deviation of a Gaussian random variable.

To compute the detection statistic,  $Z(n)$ , 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  $i = 1, \dots, M$ ), 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:

$$Z(n) = \frac{\mathbb{N}(n)}{\sqrt{\mathbb{D}(n)}} = \frac{\sum_{i=1}^M 2^{-\min(i,M-1)} [x_i(n)/\hat{\sigma}_i^2(n)] * s_i(-n)}{\sqrt{\sum_{i=1}^M 2^{-\min(i,M-1)} \hat{\sigma}_i^{-2}(n) * s_i^2(-n)}}. \quad (9.6)$$

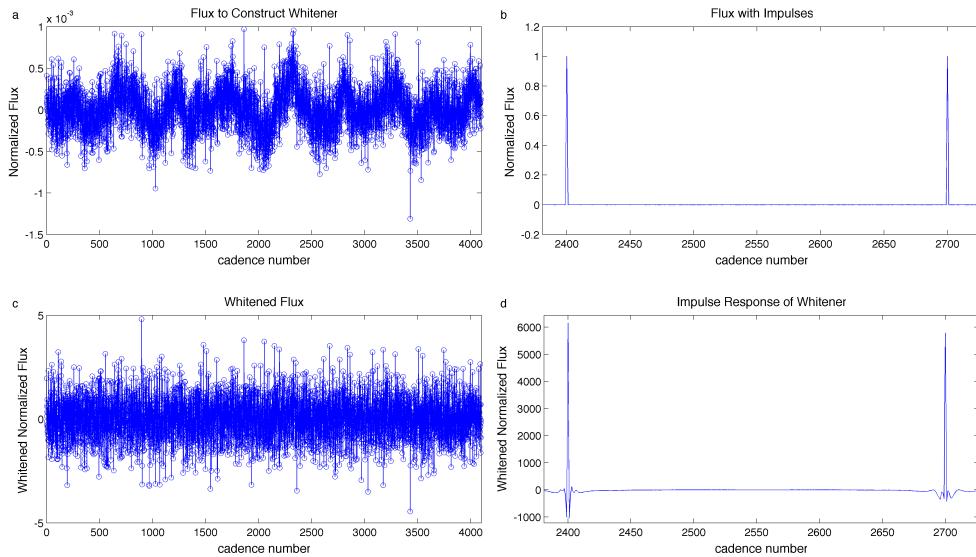
Note that the “ $-$ ” in  $s_i(-n)$  indicates time reversal. The numerator term,  $\mathbb{N}(n)$ , is essentially the correlation of the reference transit pulse with the data. If the data were WGN then the

result could be obtained by simply convolving the transit pulse with the flux time series. The expected value of Equation 9.6 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 SNR of the reference transit in the data as a function of time. The CDPP estimate is obtained as

$$CDPP(n) = 1 \times 10^6 / \sqrt{\mathbb{D}(n)}, \quad (9.7)$$

in units of parts per million.

Figure 9.6 shows the result of whitening a light curve for a star that exhibits obvious stellar variability and likely star spot modulation. Two impulses have been added to the light curve to illustrate the adaptive nature of the whitening filter.

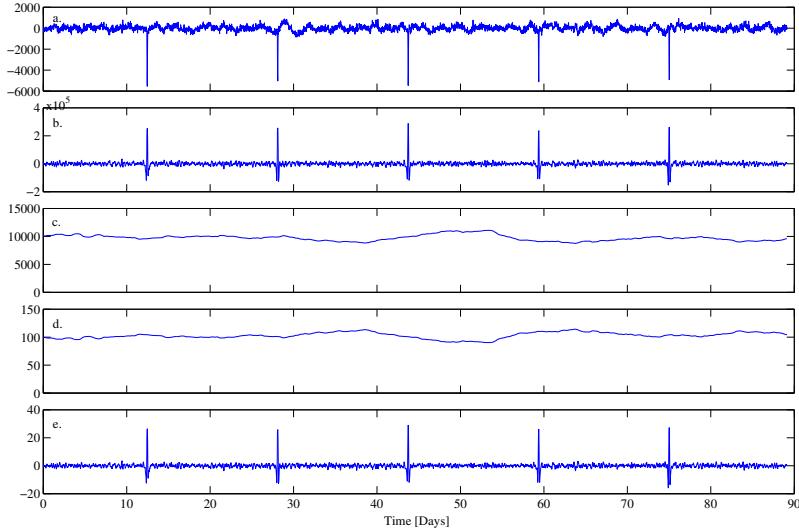


**Figure 9.6** Whitening the flux time series of a variable star. a: Normalized target flux in parts per million (ppm). b: Impulses. c: Whitened flux time series. d: Whitened impulses. Note that the waveforms are different, illustrating that the whitening filter is adaptive.

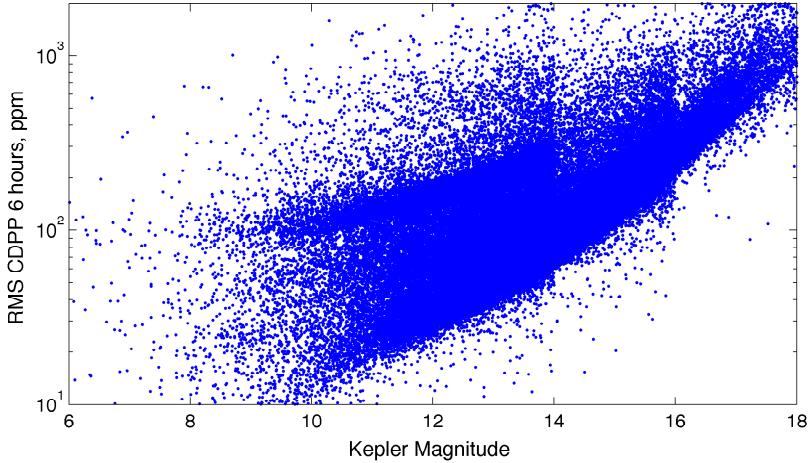
For stars with identified giant planet transits or eclipses, an alternate route is taken to estimate the correlation and expected SNR. The “in transit” cadences are removed and filled by a simple linear interpolation. The resulting time series is high-pass filtered to remove trends on timescales  $>3$  days and then a simple matched filter is convolved with the resulting time series. A moving variance (MAD) estimate supplies the information necessary to inform the expected SNR. Figure 9.7 illustrates the process of estimating CDPP for a star exhibiting strong transit-like features. Figure 9.8 shows a scatter plot of the 6-hour rms CDPP for  $\sim 200,000$  stars searched by TPS for DR25 (Twicken et al., 2016).

### 9.3.2 Removal of Positive Flux Outliers

As described in Section 2.4 of Tenenbaum et al. (2013), removal of negative flux outliers is a hazardous action, since it relies upon an algorithmic capability to distinguish between a true outlier and a transit, and for obvious reasons removing the latter is frowned upon. For this reason, strict limitations are placed upon the algorithm’s capabilities for removing suspected negative outliers.



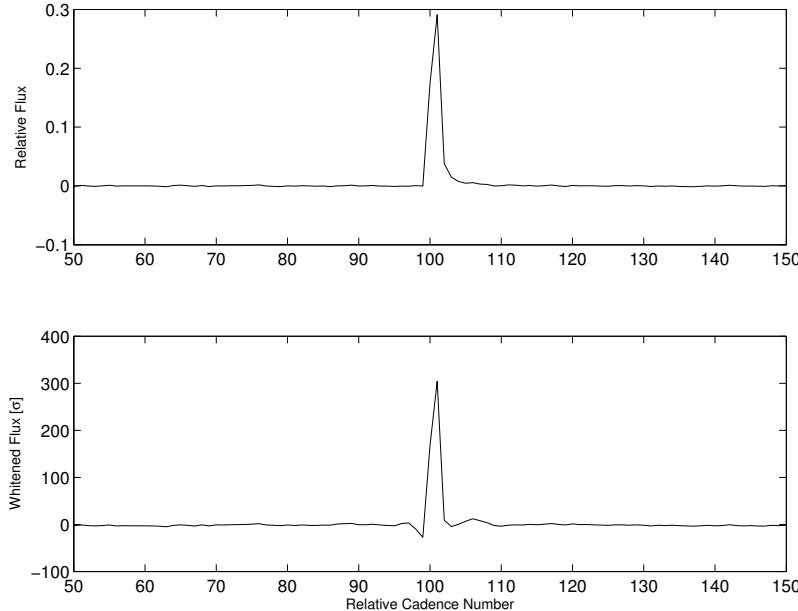
**Figure 9.7** Time series outputs of TPS for one target star. a: Normalized target flux in parts per million (ppm). b: Correlation time series  $\mathbb{N}(n)$  from numerator term of Equation 9.6. c: Normalization time series  $\mathbb{D}(n)$  from denominator term of Equation 9.6. d: 3-hr CDPP time series. e: Single-event statistic time series,  $Z(n)$ . In all cases, the trial transit pulse,  $s(n)$ , is a square pulse of unit depth and 3-hour duration (for this demonstration). From Figure 6 of Jenkins et al. (2010b).



**Figure 9.8** Six-hour CDPP as a function of *Kepler* magnitude for 197,434 stars with  $6 \leq K_p \leq 18$  and rms CDPP between 10 and 2000 ppm. Note that the changes in density at  $K_p = 14$  and  $K_p = 16$  reflect the methodology by which the *Kepler* planetary target stars were selected. For stars with  $K_p < 14$  there are two main populations evident: the branch at  $\sim 100$  ppm are giant stars while the stars in the lower branch are dwarf stars. There is considerable scatter in the rms CDPP between and above these main branches due to variable stars.

Positive outliers are much less risky to remove, since by definition a positive outlier looks like the opposite of a transit. At first glance, one might therefore assume that positive flux outliers are irrelevant as a source of false alarms or other difficulties, since the difference between a short-duration positive flux excursion and a short-duration negative flux excursion is intuitively obvious to the most casual observer. In actuality, however, positive flux excursions can result in false

alarm detections via the following mechanism: when a positive flux excursion is subjected to the whitening filter, the whitened result includes "ringing" that precedes and follows the excursion, as shown in Figure 9.9. The strongest components of the ring-down have the opposite sign to the original excursion, thus a positive excursion in the flux results in two negative excursions in the whitened flux, which are often misconstrued as transits by the subsequent search.



**Figure 9.9** Effect of a positive flux outlier. Top: Original flux. Bottom: Whitened flux. Note that whitening introduces a negative outlier to the whitened flux, which can be misconstrued as a transit. While the resulting negative outlier is much smaller than the original positive outlier, in this case the negative outlier still has a single event significance of over  $27\sigma$ .

The removal of positive outliers is accomplished by marking their locations in the quarter-stitched flux time series as gaps and applying the standard TPS gap-filling algorithm. The identification of positive outliers, by contrast, makes use of the whitened flux. The advantage to this is that by design the whitened flux contains approximately Gaussian-distributed, zero mean, unit variance white noise, plus quasi-impulsive outliers; consequently, the positive outliers are extremely easy to identify in the whitened flux. The disadvantage is that a positive outlier in the whitened flux can either indicate a positive outlier in the original flux, or it can be part of the ring-down of a negative outlier such as a transit; this can be visualized by inverting the lower plot in Figure 9.9. Thus the algorithm for positive outlier removal is as follows:

- Whiten the quarter-stitched flux;
- Identify clusters of whitened flux values which exceed a threshold: in this case a threshold of  $12.3\sigma$  is used, as explained in Appendix B;
- determine whether each cluster is due to positive outliers in the original flux or due to the ring-down of negative outliers in the original flux, which is accomplished by examining the local minima adjacent to each cluster, since for positive outliers the local minima will be weaker than the positive outliers, whereas for the ring-down of a transit one of the local minima will be much stronger than the positive outliers;

- For each positive outlier value thus identified, mark the cadences in the quarter-stitched flux as gapped and apply gap-filling;
- Produce a new whitened flux from the outlier-removed quarter-stitched flux and iterate the process until no further positive outliers are identified, which takes into account that removal of outliers can change the local noise characteristics slightly, causing values that had previously been below threshold to exceed the threshold.

### 9.3.3 Quarter-by-Quarter Whitening

Perhaps the most important change to TPS in SOC 9.3 is the introduction of quarter-by-quarter whitening. The previous codebases stitched all quarters of data together for each target star, then whitened the quarter-stitched flux time series and searched for transit-like features. The current version of TPS separately whitens the flux time series for each quarter prior to concatenating the quarterly segments together in preparation for the transit search. This change was motivated by the fact that a significant population of false alarms were generated near boundaries of the quarterly data segments. Inspection revealed that this was due to step changes that occurred in noise statistics and power from quarter to quarter as stars rotated onto new CCDs as a result of roll maneuvers. The adaptive wavelet-based matched filter algorithm in TPS was designed to track gradual changes in the statistics of observation noise, not abrupt changes such as those engendered by roll maneuvers. Quarter-by-quarter whitening significantly reduces the incidence of false alarms near quarterly boundaries.

### 9.3.4 Setting De-emphasis Weights

De-emphasis weights are applied to the correlation component,  $\mathbb{N}(n)$ , of the SES time series in order to allow the transit detector to selectively ignore data inside of gaps and in the neighborhood of long gaps, as residual transients near data gaps tend to generate spurious false alarms. In SOC 9.3 the de-emphasis window length was set to begin one cadence after safe modes (as well as for gaps longer than 1.5 days) and last 2 days. The window length after attitude tweaks was set to begin one cadence after the tweak and last for 12 cadences ( $\sim 6$  hours). First, the de-emphasis parameter,  $p$ , was set for each cadence so that it ramped linearly from 0 at the edge of the gap to 1 at the boundary of the window size. Then the de-emphasis weights,  $w_{deemphasis}$ , were set according to

$$w_{deemphasis} = p + (1 - p)(1 - e^{-2p}). \quad (9.8)$$

This produces a de-emphasis weighting profile that is 1 at the outer edge of the de-emphasis window farthest from the gap with a slope of approximately 1 at that edge and that then drops exponentially by two  $e$ -foldings to 0 at the edge of the de-emphasis window adjacent to the gapped region.

Once the de-emphasis weights are set, the actual transit search can commence.

## 9.4 Folding the Detection Statistics and Applying Veto

The third and final stage of TPS is to fold the single-event detection statistics developed in Sub-section 9.3.1 over the range of potential orbital periods and to apply the vetoes to those signatures that cross the  $7.1\sigma$  MES threshold. Due to the large number of false alarms from residual instrumental effects, this has become an iterative process, as the strongest MES may not pass one or more of the additional vetoes. In these cases, the period and epochs that contributed to the failed TCE are notched out of the MES parameter space and the other phases of the trial orbital period

yielding the original potential TCE are examined for MES values above the  $7.1\sigma$  threshold. If none are found, the redacted single event time series are refolded over all trial orbital periods to allow TPS to identify and inspect MES values at other orbital periods for potential TCEs. In addition, TPS has the opportunity to remove up to two strong individual features in the light curve that produce high MES values without folding onto comparably strong features to generate the high MES value.

These processing steps are described in the following subsections: folding the single event statistics is discussed in Subsection 9.4.1. The waveforms chosen for the transit pulse templates and their spacing are discussed in Subsection 9.4.2. Subsection 9.4.3 describes the limits imposed on the transit duty cycles to reduce the size of the parameter space to be searched. The  $\chi^2$  vetoes are introduced and developed in Subsection 9.4.4. Finally, removal of non-periodic transit-like features is described in Subsection 9.4.5.

#### 9.4.1 Folding the Single Event Statistics

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 cannot 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 Jenkins et al. (1996), one measure of sensitivity is the correlation coefficient between the model planetary signatures of neighboring points in parameter space. The minimum correlation coefficient,  $\rho$ , required between neighboring models, determines 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,  $\Delta t_0$ , 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,  $\rho$ , we have  $\Delta t_0 = (1 - \rho)D$ . The step size in orbital period,  $\Delta T_p$ , is strongly influenced by the number of transits expected in the dataset 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 dataset to the trial period, so that:

$$\Delta T_p = 4(1 - \rho)D/N = 4\Delta t_0/N. \quad (9.9)$$

The value used in SOC 9.3 was  $\rho = 0.95$  for orbital period and epoch. The minimum and maximum periods searched by TPS are additionally limited by astrophysical considerations, as discussed in Subsection 9.4.3.

Trial transit duration uses a SOC 9.3 value of  $\rho = 0.8$  for the transit duration minimum correlation coefficient. The trial transit durations explicitly searched by TPS are approximately 1.5, 2, 2.5, 3, 3.5, 4.5, 5, 6, 7.5, 9, 10, 12, 12.5, and 15 hours.<sup>6</sup> The spacings are not regular, as TPS is required to furnish CDPP metrics for all stars at 3, 6, and 12 hours.

Starting with the minimum trial orbital period (nominally 0.5 days), TPS applies Equation 9.9 to determine the next trial orbital period, continuing until the maximum trial orbital period, half the length of the time series, is reached. To form a multiple-event statistic for a 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 into each

<sup>6</sup>These durations are approximate because they are set in cadences, which are 29.4 minutes long, so the shorted duration searched is three LC, or 1.47 hours.

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 McCauliff et al. (2010). Figure 9.10 illustrates this process for the flux time series appearing in Figure 9.7.

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 long cadence 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 square transit pulse lasts six long cadence samples:  $[0, -1, -1, -1, -1, -1, -1, 0]$ . Shifting this transit by 9.8 minutes ( $1/3$  LC) earlier we obtain the sequence  $[-\frac{1}{3}, -1, -1, -1, -1, -1, -\frac{2}{3}, 0]$  with corresponding single-event detection statistics  $\mathbb{N}_{+1/3}(n)$  and  $\mathbb{D}_{+1/3}(n)$ . Shifting the original transit pulse by 9.8 minutes later, we obtain the sequence  $[0, -\frac{2}{3}, -1, -1, -1, -1, -1, -\frac{1}{3}]$  with corresponding single-event detection statistics  $\mathbb{N}_{-1/3}(n)$  and  $\mathbb{D}_{-1/3}(n)$ . The results are combined from all three analyses schematically as

$$\mathbb{N}(n) = \{\dots, \mathbb{N}_{+1/3}(k), \mathbb{N}_0(k), \mathbb{N}_{-1/3}(k), \mathbb{N}_{+1/3}(k+1), \mathbb{N}_0(k+1), \mathbb{N}_{-1/3}(k+1), \dots\}, \quad (9.10)$$

where the original time series is denoted by  $\mathbb{N}_0(n)$ . 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.

To improve the sensitivity of the search, astrophysics-motivated transit waveform templates were adopted, as described in the next section.

#### 9.4.2 Search Templates and Template Spacing

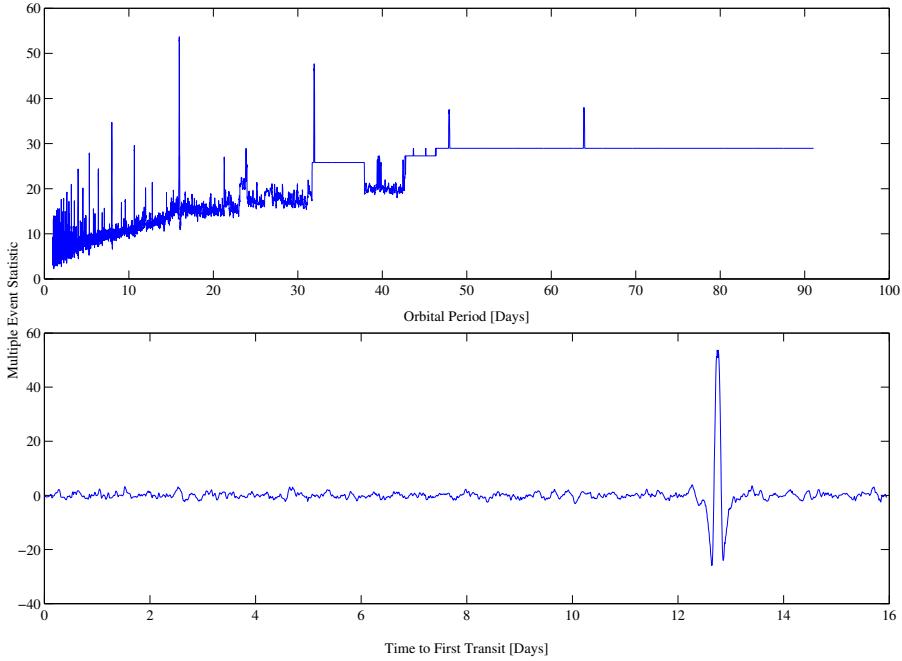
Mismatch between any true signal and the template used to filter the data degrades the SNR. This degradation due to signal-template mismatch can be decomposed into two separate types: shape mismatch and timing mismatch. The transit duration and assumed transit model affect the shape mismatch, while the template search grid spacing in the orbital period and epoch affect the timing mismatch.

Prior to the first transit search of all 17 quarters of *Kepler* data (Seader et al., 2015), TPS had simply used a square-wave transit model. In Seader et al. (2013), it was shown through a Monte Carlo study, that with perfect duration and timing match, the square wave, on average, mismatches a true signal by 3.91%. This translates directly into SNR loss. This same Monte Carlo study was used to compute the integral average of all astrophysical models based on the Mandel and Agol geometric transit model (Mandel & Agol, 2002) with limb darkening of Claret (Claret & Bloemen, 2011), over the parameter space of interest (Seader et al., 2013). TPS now uses this averaged model to construct templates, which lowers the shape mismatch to only 1.49%. Lowering this shape mismatch also improves the sensitivity of the  $\chi^2$  vetoes (discussed in the next section) which previously assumed a perfect match between the signal and template. In SOC 9.3 the calculation of the  $\chi^2$  vetoes now takes into account the signal-template mismatch as described first in Allen (2005) and later in Seader et al. (2013).

#### 9.4.3 Limitation on Allowable Transit Duty Cycles

An additional means of separating likely transiting planet signatures from false alarms is to apply bounds to the ratio of the transit duration,  $\tau$ , to the orbital period,  $T$ . The relationship between these parameters of the transit can be derived from Kepler's laws of motion as:

$$\tau = k T^{1/3}. \quad (9.11)$$



**Figure 9.10** Multiple-event statistics (MES) 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 9.7. Bottom: MES as a function of lag time for a 15.97-day period, showing a peak at 12.74 days, corresponding to the mid-time of the first transit shown in Figure 9.7. From Figure 8 of Jenkins et al. (2010b)

Equation 9.11 can be rewritten in terms of the transit duty cycle  $\phi_{\text{dut}} \equiv \tau/T$ :

$$\phi_{\text{dut}} = k T^{-2/3}. \quad (9.12)$$

The value of  $k$  for a specific system is a function of the star's properties, and also the eccentricity of the orbit under consideration. For circular orbits about the Sun,  $k$  is approximately 0.058 days<sup>2/3</sup> (Gilliland et al., 2000); for a circular orbit about a late-type M dwarf star,  $k$  is approximately 0.026 days<sup>2/3</sup>.

The shortest orbital period included in TPS searches is set to 0.5 days. At this limit, Equation 9.12 shows that  $\phi_{\text{dut}}$  for a Sun-like star and a circular orbit is approximately 0.092. To allow a margin for elliptical orbits or stars far from Solar in their parameters, we limit the maximum allowed value of  $\phi_{\text{dut}}$  to 0.16. This restriction is implemented by adjusting the minimum search period for each trial transit duration used in the search: for 1.5-hour transits, the search is allowed to operate down to periods of 0.5 days, while for 15-hour transits the minimum search period is limited to 3.9 days.

An additional restriction is also placed on the lower bound of allowed  $\phi_{\text{dut}}$  values, specifically

$$\phi_{\text{dut}} \geq 0.017 T^{-2/3}, \quad (9.13)$$

where  $T$  is the orbital period in days. This limit is 3.4 times smaller than that expected for Solar stars and 1.5 times smaller than expected for late M dwarf stars, which allows margin for elliptical orbits, large impact parameter values, and non-Solar parameters. Equation 9.13 sets a

maximum search period which is a function of transit duration: for example, 1.5-hour transits are limited to search periods of 50 days or less, while 3.0-hour transits are limited to search periods of 300 days or less.

Note that the use of the duty cycle cuts creates an implicit trade-off between the purity of the search (i.e., rejection of false positives) and its efficiency (i.e., acceptance of true positives). Specifically, the cuts will reject false positive detections around Sun-like stars on the basis of unphysicality. At the same time, the upper limit on permitted duty cycle will reject some true positive detections on stars significantly larger than the Sun; the lower limit on permitted duty cycle will reject some true positive detections on stars significantly smaller than the Sun. Both cuts can potentially reject true positive detections on Sun-like stars from planets with highly eccentric orbits. It is our judgement that the regions of parameter space excluded by these cuts are acceptably small when both sides of the trade-off are considered.

#### 9.4.4 False Alarm Vetoes

During the transiting planet search, TPS steps through potential candidates across period-epoch space with MES values exceeding the search threshold of  $7.1\sigma$  in order of decreasing MES for each pulse duration and then subjects each potential TCE to a suite of three statistical vetoes. The search continues until either TPS runs out of time on that pulse duration, it hits the maximum allowable number of candidates to loop over (set to 1000), it exhausts the list without finding anything that passes all the vetoes, or it settles on something that passes all the vetoes. During the course of performing the search, TPS has the ability to remove up to two features in the data that contribute to detections that do not pass all the vetoes (Tenenbaum et al., 2013, see Subsection 9.4.5). After removing features, the period-epoch folding is re-done to generate a new list of candidates.

Once a potential TCE with  $MES > 7.1\sigma$  is found, TPS uses the candidate signature's ephemeris to de-trend and re-whiten the light curve to avoid any loss in SNR that would otherwise occur because of the effect of the signal on the estimated trend and the whitening coefficient estimates. This improves the discriminating power of these three statistical tests, or vetoes to which TPS subjects each candidate transit sequence. The ephemeris of the candidate is used to identify in-transit cadences (with some small amount of padding). These in-transit cadences are then filled using an adaptive auto-regressive gap prediction algorithm. A trend line is then estimated and removed from the data using a piecewise polynomial fitting algorithm that employs AIC to prevent over-fitting. The whitening coefficients are then re-computed. After removing the trend from the in-transit cadences, they are restored to the trend-removed data, which are then whitened using the new whitening coefficients. The re-whitened data are now subjected to the robust statistic and the two  $\chi^2$  statistics,  $\chi^2_{(2)}$  and  $\chi^2_{(GOF)}$ .

The threshold for the robust statistic test was set to  $6.8\sigma$  in the DR25 run and rejected 43,313 signatures with  $MES > 7.1\sigma$ . That is, these signals failed the robust statistic test and a lower MES signal was not subsequently identified that both met the detection threshold and passed all vetoes, and so no TCE was generated. The surviving 24,691 signatures were subjected to two different  $\chi^2$  tests (Seader et al., 2013). Thresholds for the robust statistic and  $\chi^2$  tests were tuned by analysis of TPS performance for target stars with injected transit signatures and a subset of known KOIs. The thresholds were set as high as possible to facilitate false alarm rejection without significantly impacting the recovery of true transit signals.

These statistical vetoes are described in more detail below.

**9.4.4.1 The Robust Statistic** The robust statistic (RS) performs a kernel-based robust fit of the in-transit data to the putative transit signature and normalizes the fitted transit depth by the fit uncertainty (Tenenbaum et al., 2014). This test penalizes outliers that erroneously contribute to a high detection statistic (MES). There is an additional criterion applied during the RS test that

affects only candidates with the minimum allowed number of transits (three transits). We require that each transit has no more than 50% of its cadences with data quality weights less than unity (Tenenbaum et al., 2014).

The first step in constructing the RS is to generate the transit model pulse train. This consists of a train of transit pulses that are positioned at the locations of the transits as determined by the period and epoch associated with the MES. Let  $\mathbf{s}$  be this model pulse train vector. The pulse train  $\mathbf{s}$  and the data, or flux time series,  $\mathbf{x}$  are each whitened to eliminate the effect of stellar variations. The whitened model and data vectors,  $\tilde{\mathbf{s}}$  and  $\tilde{\mathbf{x}}$  (where ‘ $\sim$ ’ denotes a whitened vector), are then windowed to remove out-of-transit samples. The resulting whitened and windowed transit model  $\tilde{\mathbf{s}}$  is then robustly fit to the whitened and windowed data  $\tilde{\mathbf{x}}$  to generate a diagonal matrix of fit weights  $\mathbf{W}$ . The RS,  $Z_{\text{RS}}$ , is then calculated as:

$$Z_{\text{RS}} = \frac{\tilde{\mathbf{s}}^T \mathbf{W} \tilde{\mathbf{x}}}{\sqrt{\tilde{\mathbf{s}}^T \mathbf{W} \tilde{\mathbf{s}}}}, \quad (9.14)$$

where  $T$  denotes the transpose of a vector, and where Equation 9.14 is applied only to data samples within the transit windows described above.

In the limit in which the data vector  $\mathbf{x}$  and the model vector  $\mathbf{s}$  are well-matched in shape and duration, the matrix  $\mathbf{W}$  will approach the identity matrix and  $Z_{\text{RS}}$ , as defined in Equation 9.14, will be approximately equal to the MES. In reality, the match between data and model is imperfect: the transits in the model vector are represented as approximate transit pulses rather than true transit shapes, and in general the duration of the trial transit pulse and the true transits will not be identically matched to one another. In studies of known transiting planet systems, this mismatch can lower  $Z_{\text{RS}}$  by about 10% compared to the MES ( $Z$ ).

Now consider a situation in which the MES is constructed from folding a single, extremely strong transit-like signature over two or more events consistent with statistical fluctuations, which is a typical case of non-uniform-depth events being combined into a MES which lies above threshold. Because the fit is performed robustly, the weak transit-like signatures will “out-vote” the strong one, leading to near-unity weights for the weak events and near-zero weights for the strong event. When the weights in this instance are combined with the data and model vectors as shown in Equation 9.14, the result will be a low value for  $Z_{\text{RS}}$ . It is in this way that the RS permits events with significant transit depth mismatches to be vetoed while preserving events with relatively uniform transit depths.

**9.4.4.2 The  $\chi^2_{(2)}$  Statistic** The  $\chi^2_{(2)}$  test breaks up the MES into different components, one for each transit event, and compares what is expected from each transit to what is actually obtained in the data, assuming that there is indeed a transiting planet (Seader et al., 2013, 2015). We begin by examining the sum of the contributions of each transit to the detection statistic,  $Z$ , defined in Equation 9.6 for a specific orbital period, epoch, and transit duration in terms of the whitened data,  $\tilde{\mathbf{x}}$ , and whitened signal template,  $\tilde{\mathbf{s}}$ :

$$Z = \sum_{j=1}^P z_j, \quad (9.15)$$

where  $P$  is the number of transits and

$$z_j = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{s}}_j}{\sqrt{\tilde{\mathbf{s}}_j \cdot \tilde{\mathbf{s}}_j}}, \text{ for } j = 1, \dots, P, \quad (9.16)$$

where  $\tilde{\mathbf{x}}$  is the whitened data vector and  $\tilde{\mathbf{s}}_j$  is the  $j^{\text{th}}$  whitened transit pulse vector. Define the quantity,  $q_j$ , representing the fractional expected contribution of each transit pulse to  $Z$ , as

$$q_j = \frac{\tilde{\mathbf{s}}_j \cdot \tilde{\mathbf{s}}_j}{\sum_i \tilde{\mathbf{s}}_i \cdot \tilde{\mathbf{s}}_i}, \text{ for } j = 1, \dots, P. \quad (9.17)$$

The sum of the  $q_j$  terms over all transits in Equation 9.17 is identically 1:

$$\sum_{j=1}^N q_j = 1. \quad (9.18)$$

The difference,  $\Delta z_j$ , between the observed contribution to the MES and the expected contribution for each transit pulse is

$$\Delta z_j = z_j - q_j Z, \quad (9.19)$$

for  $j = 1, \dots, P$ , and it can be shown that

$$\sum_{j=1}^P \Delta z_j = 0. \quad (9.20)$$

The  $\chi^2_{(2)}$  statistic is defined as

$$\chi^2_{(2)} = \sum_{j=1}^P \frac{(\Delta z_j)^2}{q_j}, \quad (9.21)$$

and has  $\langle \chi^2_{(2)} \rangle = P - 1$  degrees of freedom.

In SOC 9.3, the calculation of the number of degrees of freedom for the  $\chi^2_{(2)}$  veto was updated to account for the mismatch between the non-rectangular, astrophysics motivated transit pulse templates and actual limb-darkened transit pulses. Monte Carlo experiments using artificial transit injection were conducted to determine an empirical correction factor  $\epsilon$  to the number of degrees of freedom for this veto as

$$\langle \chi^2_{(2)} \rangle = n_{Transits} - 1 + \frac{\epsilon (2 - \epsilon)}{(1 - \epsilon)^2} Z^2, \quad (9.22)$$

where  $\epsilon \approx 0.04$ . This value was also supported by analysis of the set of “golden KOIs” against which the performance of TPS and DV are routinely measured (see, e.g., Seader et al., 2015; Twicken et al., 2016, Subsection 9.5.4). The empirical correction factor was initially formulated by Allen (2005).

The reduced  $\chi$ -square, the ratio of the value of  $\chi^2_{(2)}$  to its expected value,  $\langle \chi^2_{(2)} \rangle$ , provides a measure of the degree to which the contributions of the individual transit pulses to the total detection statistic are comparable, and can thus be used to obtain a modified MES that is thresholded (against  $7\sigma$  for DR25):

$$Z_{(2)} = Z \left( \frac{\chi^2_{(2)}}{\langle \chi^2_{(2)} \rangle} \right)^{-1/2}. \quad (9.23)$$

**9.4.4.3 The  $\chi^2_{(GOF)}$  “Goodness of Fit” Veto** This veto measures the strength of the residual errors of the transit signature fit to the data. This can also be interpreted as the difference between the squared amplitude of the detector output and the squared SNR (Bagglio et al., 2000; Allen, 2005). There are several subtleties, described in Seader et al. (2013), associated with the construction of  $\chi^2_{(2)}$ . These subtleties must also be taken care of in the construction of this  $\chi^2_{(GOF)}$  statistic, and in what follows it is assumed that they are. The mathematical development of  $\chi^2_{(GOF)}$  and  $\chi^2_{(2)}$  are similar but they partition the information at different levels:  $\chi^2_{(2)}$  concerns itself with the strength of the contributions of the individual transit pulses to the total SNR, while  $\chi^2_{(GOF)}$  measures the contributions of the individual cadences.

As developed in Appendix B of Seader et al. (2015),  $\chi^2_{(GOF)}$  can be expressed in terms of the whitened data vector,  $\tilde{x}$ , and the whitened transit signature,  $\tilde{s}$ :

$$\chi^2_{(GOF)} = \sum_{n=1}^N \tilde{x}(n)\tilde{x}(n) - \frac{\left[\sum_{n=1}^N \tilde{x}(n)\tilde{s}(n)\right]^2}{\sum_{n=1}^N \tilde{s}^2(n)} \quad (9.24)$$

$$= \tilde{x} \cdot \tilde{x} - \frac{(\tilde{x} \cdot \tilde{s})^2}{|\tilde{s}|^2} \quad (9.25)$$

which has  $\langle \chi^2_{(GOF)} \rangle = N_t$  degrees of freedom, where  $N_t$  is the number of in-transit cadences in the time series. Note that  $x_j(n)$  is zero outside of transit  $j$  as discussed in Seader et al. (2013). As with the  $\chi^2_{(2)}$  veto, TPS thresholds the result of normalizing the MES by the reduced  $\chi_{(GOF)}$  statistic:

$$Z_{(GOF)} = Z \left( \frac{\chi^2_{(GOF)}}{\langle \chi^2_{(GOF)} \rangle} \right)^{-1/2}. \quad (9.26)$$

The threshold for  $Z_{(GOF)}$  was set to  $6.8\sigma$  for DR25 (decreased from  $7.0\sigma$  in the DR24 run).

The two *chi*-square vetoes complement the MES and the RS, which measure the correlation between the data and the fitted waveform. In contrast,  $\chi^2_{(2)}$  and  $\chi^2_{(GOF)}$  measure the degree to which the fit residuals meet expectations, or the degree to which the scatter of the differences between the data and the fitted model is consistent with a zero-mean, unit variance noise process as expected in the whitened domain. A total of 7461 of the potential DR25 TCEs that survived the RS test failed at least one of the  $\chi^2$  tests and 5147 failed both tests; these represent the  $\chi^2$  test failures after which a lower MES TCE was not subsequently identified. The number of potential TCEs rejected only by the  $\chi^2_{(2)}$  test was 1,990 while the number rejected only by the  $\chi^2_{(GOF)}$  test was 324. A total of 17,230 signatures with  $MES > 7.1\sigma$  survived all statistical checks and were processed through Data Validation; this includes 1779 TCEs identified after stronger MES signals were vetoed.

#### 9.4.5 Removal of Non-Periodic Transit-Like Features

The benefits of the multiple iterations of search, described above, can only be fully exploited in the absence of relatively strong non-astrophysical single events. Such strong events will cause the MES to exceed the  $7.1\sigma$  threshold for large numbers of possible periods: folding a single strong event with a small number of weak events will produce a large MES, and there are an extremely large number of period-epoch combinations that will result in such a folding. If this happens, the 1000 iterations of searching can easily be exhausted in the process of eliminating a fraction of the spurious MES caused by a single strong event. Such an outcome can be avoided if these strong events are identified and removed prior to folding, but such removals are obviously dangerous: without prior knowledge, a feature in the data that is identified as a non-astrophysical event and removed could actually be a strong transit. For this reason, any event removal must be used sparingly. TPS addresses this issue in two ways. First, a minimum number of transits is required for an event to be accepted, since the probability of such chance combinations yielding a MES over the threshold decreases as the number of events folded together increases. At present, the threshold number of transits is three. Second, the current version of TPS is permitted to remove one, and only one, single event, and only in the case in which the first iteration of planet searching produces a strongest MES that exceeds the MES threshold of  $7.1\sigma$  but that is then vetoed by RS,  $\chi^2_{(2)}$ , or  $\chi^2_{(GOF)}$ . In such a case, the strongest single event in the time series is removed, if and only if the strongest single event has an amplitude that is greater than the MES threshold multiplied by the square root of the minimum number of transits ( $7.1\sigma \times \sqrt{3}$ , or  $12.3\sigma$  for the current parameter choices).

## 9.5 Performance of TPS in the DR25 Search

The SOC 9.3 transiting planet search of all 17 quarters of *Kepler* data included a total of 198,709 stellar targets of which 112,046 were observed in all 17 quarters and 86,663 in fewer than 17 quarters. There were 17,230 targets for which at least one transit signature was identified that met the specified detection criteria: periodicity, minimum of three observed transit events, detection statistic in excess of the search threshold, and passing grade on three statistical transit consistency tests. Light curves for which a transit signal was identified were iteratively searched for additional signatures after a limb-darkened transiting planet model was fitted to the data and in-transit data were removed. The search for additional planets added 16,802 transit signals for a total of 34,032; this far exceeds the number of transit signatures identified in prior pipeline runs. There was a strategic emphasis on completeness over reliability for the final *Kepler* transit search. The recovery rate against a set of 3,402 well established, high quality Kepler Objects of Interest (KOI) yielded a recovery rate of 99.8%. These results are more fully documented by Twicken et al. (2016).

### 9.5.1 Incomplete Searches

*Kepler* Pipeline modules are subject to total processing time, or “wall time,” limits when run on the NAS Pleiades supercomputer. Computational tasks that are still running on Pleiades when the wall time limit is reached are killed; it is imperative that pipeline tasks not suffer this fate because results are not produced in such eventualities as the processes are simply “killed”. To prevent the wall time being exceeded, time limits are allocated to subcomponents of TPS with margins against the total time limit that are monitored during the processing. The time limits for TPS and DV must accommodate transit searches and transit consistency checks, transiting planet model fitting, computation of diagnostic metrics, and generation of reports. The time limits are managed per target by enforcing limits on the number of TCEs, transiting planet search and consistency check iterations, and model fit and diagnostic test iterations. Furthermore, self-timeouts are enforced on a number of TPS and DV processing steps. It should be noted that whereas all TCEs produced in the pipeline are guaranteed to meet the pipeline detection threshold, there is no guarantee that a TCE represents the “best” detection (i.e., highest possible MES with a passing grade on all consistency tests) for the given light curve in the gridded search space. A TCE may represent the best result that could be achieved in the time available.

In the DR25 run, 35 hours were allocated for each TPS work unit compared to 50 hours for the DR24 run. The decision to do so was based on non-technical considerations. Decreasing the time limit resulted in a small population of targets (1.3% of the total) for which not all trial transit pulse durations were searched for transits<sup>7</sup>.

Of 2497 targets that timed out before visiting all pulse durations, 1952 generated TCEs nonetheless. Indeed, only 545 of the targets that timed out failed to generate a TCE. We can determine whether significant sensitivity was lost due to the change in TPS time limit by investigating the fates of the 587 known KOIs for which not all pulse durations were searched in the DR25 run. Of these KOIs, 12 are confirmed or validated planets, 137 are dispositioned as planet candidates (PC), and 438 are false positives (FP). TCEs were generated for 562 of these KOIs, most at the expected ephemeris. All 12 confirmed or validated that planet light curves generated TCEs at the correct period and epoch. Only 25 of the 587 known KOIs failed to produce TCEs: 4 of these were PCs, and 21 were FPs.

<sup>7</sup>TPS searches each light curve for transits over 14 distinct pulse durations from 1.5 hours to 15 hours, starting from the longest pulse duration and working toward the shortest (as longer period planets are of greater interest than shorter period planets). If TPS reaches an internal timeout before the shortest pulse durations are searched, it will abort the search to preserve the results from the other pulse durations.

We conclude the following for the PCs that failed to produce TCEs based on DR24 transit search results: 1) KOI 6262.01 is consistent with a transiting planet with orbital period (0.34 days) below the minimum searched in the pipeline (0.5 days); 2) KOI 7572.01 features purported transits with a period of 91.1 days that were largely gapped<sup>8</sup> by two short-period TCEs on the same target and does not appear to be a credible planet candidate; 3) KOI 6918.01 appears likely to be an eclipsing binary and has been classified as such by Kirk et al. (2016); 4) KOI 6598.01 features purported transits with a period of 11.0 days that were largely gapped by three short-period TCEs triggered by strong stellar variability and also does not appear to be a credible planet candidate. Thus, the self-timeouts affected none of the confirmed or validated planets, few credible planet candidates, and a small number of astrophysical false positives. While unfortunate, the incomplete transit searches in TPS are unlikely to significantly impact catalog completeness.

The search for additional planets in light curves with TCEs is conducted in DV (through internal calls to the main TPS function); the time allocated to these searches is also subject to run time constraints. Sufficient processing time must be held in reserve so that the DV diagnostic tests may be performed and reports generated for all TCEs identified on a given target. The time limit on transit searches for additional planets in light curves with TCEs depends upon the total DV time allocation for each work unit (which did not change from DR24 to DR25) rather than the TPS allocation per work unit (which was reduced from DR24 to DR25). In the limit, the multiple planet search is halted before any of the 14 trial pulse durations are searched if insufficient time would then be available to complete Data Validation.

### 9.5.2 Detection of Multiple-Planet Systems

For the 17,230 target stars found to contain a TCE, additional transit searches were employed to identify potential multiple-planet systems. The process is described in Wu et al. (2010); Tenenbaum et al. (2013), and in Chapter 11. The multiple-planet search incorporates a configurable upper limit on the number of TCEs per target, which is currently set to 10. This limit was established for two reasons. First, the limit on TCEs for a given target was instituted to manage pipeline task processing time as described earlier. Second, applying a limit to the number of TCEs per target prevents a failure mode in which a target flux time series is sufficiently pathological that the search process becomes “stuck,” returning one detection after another. The selected limit of 10 TCEs is based on experience: the maximum number of KOIs to date on a single target star is seven, which indicates that limiting the process to ten TCEs per target is unlikely to sacrifice potential KOIs.

The transit searches performed for multiple planet systems yielded 16,802 additional TCEs across 7120 unique target stars, for a total of 34,032 TCEs. The number of TCEs identified on each target in the DR25 transit search is listed in Table 1 of Twicken et al. (2016) and is available online in machine readable format. The average number of TCEs per target with at least one potential transit signal is 2.0. Figure 9.11 shows the distribution of the number of targets with each of the allowed numbers of TCEs; the number of targets is displayed on a logarithmic scale. The DR25 results represent the largest number of TCEs that have ever been produced in a pipeline run for generation of a catalog of planetary candidates. The *Kepler* Pipeline development team was encouraged by the scientific community to increase sensitivity to small planets in long-period orbits and to emphasize completeness over reliability. Accordingly, a large number of the

<sup>8</sup>Flux data on cadences in and near transit for a given TCE are removed (“gapped”) from the flux time series in DV before the search is invoked for additional transit signatures. The transits of subsequent, generally lower MES signatures that overlap the events associated with prior TCEs on the same target are therefore unavailable for the transit search algorithm and do not contribute to the detection MES. Short-period TCEs leave a periodic train of short data gaps in the residual flux time series searched for subsequent transit signals; these time series have been referred to as “Swiss cheese.” TCEs identified subsequent to short-period TCEs on a given target should be closely examined for validity.

34,032 TCEs are false alarm detections. A variety of approaches may be undertaken to identify the subset of legitimate planet candidates from the large population of TCEs.

For the record, there were 7492 total TCEs in 1,006 systems with six or more TCEs. The limit of 10 TCEs was reached for 139 targets. A small number of the TCEs in systems with six or more TCEs represent bona fide planet candidates and in some cases confirmed planets (e.g., KOI 157/*Kepler*-11 system, KOI 435/*Kepler*-154 system, KOI 351/*Kepler*-90 system). The vast majority of these TCEs are false alarms, however. There are not nearly this many detectable transiting planets in large multiple-planet systems in the *Kepler* dataset. The bulk of these TCEs are triggered by artifacts and other features in individual light curves that produce multiple false alarms.

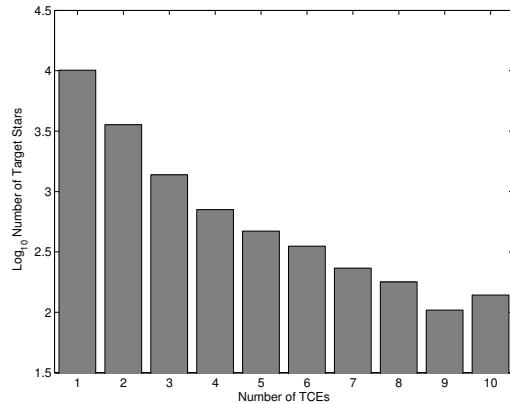
All TCEs included in this analysis have been delivered to the NASA Exoplanet Archive<sup>9</sup> along with comprehensive DV Reports for each target with at least one TCE and one-page DV Report Summaries for each TCE. The Reports and Summaries are available to Exoplanet Archive visitors in PDF format. Tabulated DV model fit and diagnostic test results are also available at the Exoplanet Archive, as are newly redesigned data products in Flexible Image Transport System (FITS) format that include TPS and DV time series data relevant to the TCEs identified in the pipeline (Thompson et al., 2016).

### 9.5.3 TCE Population

We now summarize the population of TCEs produced in the *Kepler* Data Processing Pipeline in the Q1–Q17 DR25 run and draw comparisons with the DR24 results reported by Seader et al. (2015).

The final search (DR25) for transiting planets in the full primary *Kepler* Mission dataset produced 34,032 TCEs on 17,230 unique stellar targets. This compares with 16,285 TCEs on 9743 unique targets in the Q1–Q16 search reported by Tenenbaum et al. (2014), and 20,367 TCEs on 12,669 unique targets in the Q1–Q17 DR24 search reported by Seader et al. (2015). The increase in the number of potential transit signatures is due largely to: 1) improvements in pipeline pixel calibration, photometry, and transiting planet search algorithms, 2) modifications to transit signature consistency tests and test criteria, and 3) software bug fixes. The fundamental detection threshold for the transiting planet search has not changed, however.

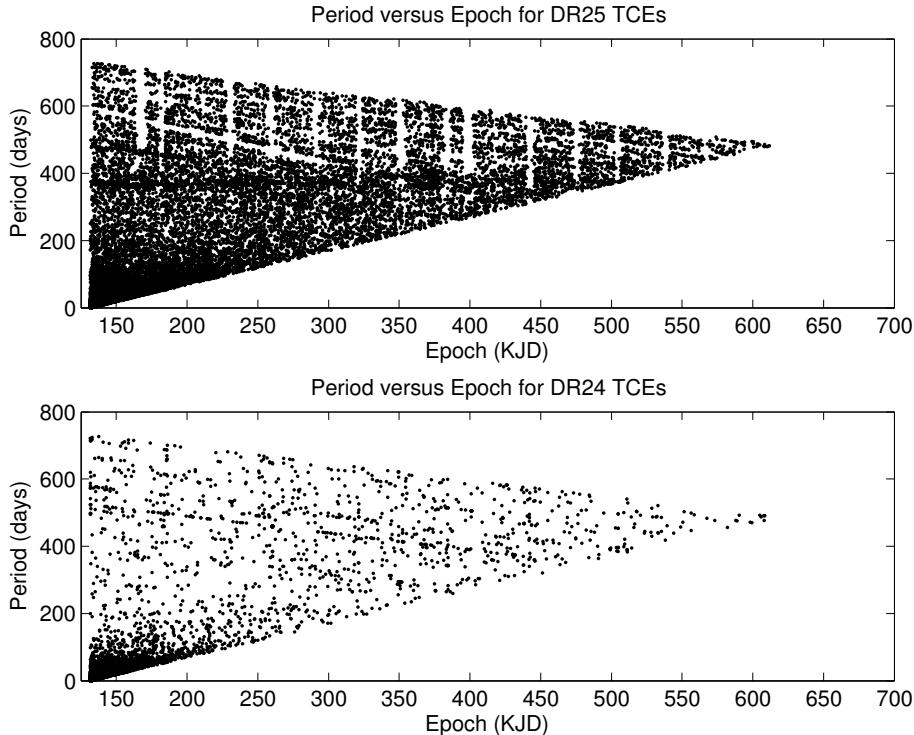
Figure 9.12 (top panel) shows the period and epoch of first transit for each of the 34,032 DR25 TCEs, with period in units of days and epoch in *Kepler*-modified Julian Date (KJD), which is Julian Date - 2454833.0 (1 January 2009). As discussed earlier in Chapter 8, structure in the ensemble of TCEs displayed in the period versus epoch “wedge” is undesirable (although white strips without TCEs are unavoidable due to gaps in the *Kepler* dataset). Figure 9.12 (lower



**Figure 9.11** Distribution of transit search targets by number of associated TCEs. The number of targets is displayed on a logarithmic scale. The maximum number of TCEs per target was configured to be 10. In total, 34,032 DR25 TCEs were identified on 17,230 unique targets for an average of 2.0 TCEs per target (for targets with TCEs). From Figure 3 of Twicken et al. (2016).

<sup>9</sup><http://exoplanetarchive.ipac.caltech.edu>.

panel) shows the same plot for the 20,367 TCEs detected in the earlier DR24 pipeline run. The axis scaling is identical for the two subplots, as is the marker size. Several features are apparent in this comparison. First, the number of TCEs is considerably larger in the Q1–Q17 DR25 analysis. Second, the number of TCEs with long periods is considerably greater in the DR25 analysis.

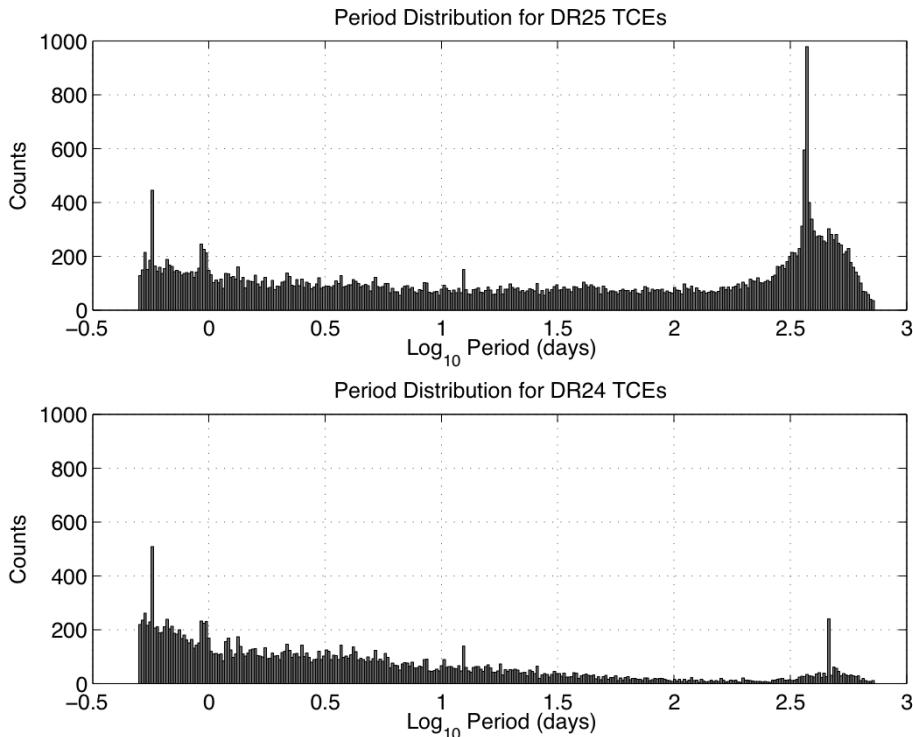


**Figure 9.12** Orbital period versus epoch “wedge” plots for the 34,032 DR25 TCEs detected in Q1–Q17 of *Kepler* data (top); and for the 20,367 DR24 TCEs detected in Q1–Q17 of *Kepler* data (bottom) as reported in Seader et al. (2015). Periods are in days, epochs are in Kepler-modified Julian Date (KJD); see text for definition. From figure 4 of Twicken et al. (2016).

The paucity of long-period TCEs in the prior analysis reported by Seader et al. (2015) may be traced largely to the TPS transit signal consistency tests that were employed in the DR24 run. In an effort to mitigate the large number of long-period false alarms due to image artifacts and the quarterly photometer roll, the TPS vetoes eliminated many long-period TCEs. Preventing false alarms while at the same time maintaining sensitivity to transit signals has always been a difficult balancing act. As discussed earlier, we have 1) improved the algorithm for computing the degrees of freedom associated with the  $\chi^2_{(2)}$  transit signal consistency test, 2) lowered the threshold on the  $\chi^2_{(GOF)}$  test, and 3) disabled the consistency test based on a statistical bootstrap. Together with improvements to the quality of the light curves and the transiting planet search algorithm, we are hopeful that a number of the long-period TCEs in this Q1–Q17 analysis represent small planets orbiting in the habitable zone of Sun-like stars even if the vast majority of the long period TCEs represent false alarms.

The drastic change in the distribution of TCE periods may be seen more clearly in Figure 9.13, which shows the distribution of TCE periods in units of days on a logarithmic scale. The Q1–Q17 DR25 results are shown in the upper panel of the figure and the Q1–Q17 DR24 results are shown in the lower panel. The distributions of TCEs by orbital period are displayed with the same axis scaling in both cases. TPS has been configured to search for transiting planet

signatures with orbital periods above 0.5 days. The minimum orbital period for the multiple planet search was reduced from 1.0 to 0.5 days approximately 1.5 years after the *Kepler* launch to address the growing population of transiting planets with periods less than 1.0 day. The minimum orbital period was never reduced further in the Pipeline TPS. There is a high computational cost involved in searching for short-period transit signatures. Short-period transiting planet signals do not necessarily go undetected. Transit signals with periods below 0.5 days often produce one or more detections at integer-multiples of the true orbital period. In cases such as these, the transit ephemerides are typically corrected by TCERT in the process of catalog generation. Failure to detect certain short period transiting planets in the pipeline is likely to be attributable to removal of harmonic content (as discussed in Subsection 9.2.2) or transit consistency vetoes (as discussed in Subsection 9.4.4) rather than a lower limit on the transit search period.



**Figure 9.13** Distribution of TCE periods plotted logarithmically. Top: 34,032 DR25 TCEs detected in Q1–Q17 of *Kepler* data. Bottom: 20,367 DR24 TCEs detected in Q1–Q17 of *Kepler* data as reported in Seader et al. (2015). The peak near 372 days ( $\log_{10} = 2.57$ ) in the DR25 results is coincident with the orbital period (and hence thermal cycle) of the *Kepler* spacecraft. The peak near 460 days ( $\log_{10} = 2.66$ ) in DR24 analysis has been eliminated by improvements to data gap filling code in the DR25 codebase. The common peaks near 0.57 days ( $\log_{10} = -0.24$ ) and 12.45 days ( $\log_{10} = 1.10$ ) are due to contamination by RR Lyrae and V380 Cyg respectively (Coughlin et al., 2014). From Figure 5 of Twicken et al. (2016).

The distribution of the latest Q1–Q17 TCEs is roughly uniform with log period up to periods on the order of 200 days. Beyond that point there is a large excess of TCEs with a distinct peak near the *Kepler* orbital period of 372 days. This excess of long-period TCEs is clearly non-astrophysical. Transit signatures at long periods are comprised of relatively few transits. Gaussian detection statistics do not necessarily apply when small numbers of transit events are folded; the false alarm rate therefore depends not only on the pipeline detection threshold but also on the number of observed transits. Furthermore, the discriminating power of the  $\chi^2$  consistency tests is diminished at long periods due to the low number of transit events and hence degrees

of freedom. The false alarm probability for each TCE is determined with a statistical bootstrap calculation in DV (see Chapter 10 and Chapter 11).

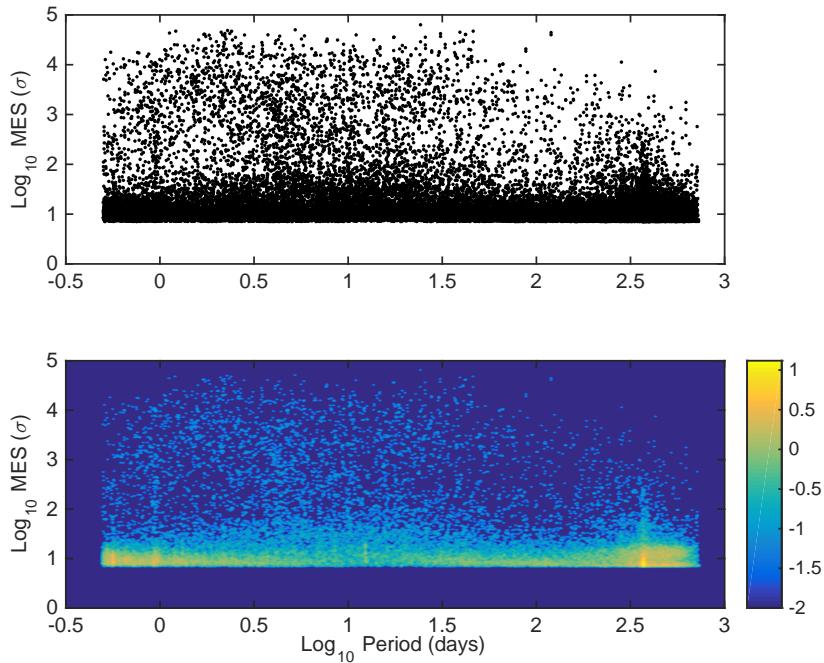
In the Q1–Q17 DR24 analysis reported by Seader et al. (2015), a bootstrap-based veto was employed in the suite of TPS transit signal consistency tests. This veto essentially enforced a detection threshold that varied by TCE to yield a uniform false alarm probability. As seen in Figure 9.13, the distribution of TCEs generally decreased with (log) period although there was also an excess at long periods. The increase in detection threshold with period reflected the non-Gaussian nature of the noise and was effective at eliminating long-period false alarms. In order to quantify the sensitivity and detection efficiency of the DR24 pipeline, transit signatures were injected at the pixel level into flight data associated with most LC targets and the *Kepler* Pipeline was subsequently run through PA, PDC, TPS and DV. In the analysis of this run by Christiansen et al. (2015), some loss in sensitivity of the DR24 pipeline to long-period transiting planets was noted. A similar transit injection activity is underway to characterize the DR25 *Kepler* Pipeline codebase (Christiansen 2017, in prep).

A close examination of Figure 9.13 reveals that there are several peaks in the TCE period histograms for the DR24/DR25 results that are highly localized. There was a long-period peak near 460 days ( $\log_{10} = 2.66$ ) in the DR24 analysis that has been eliminated by improvements to data gap filling code in SOC 9.3. There is a peak in the DR25 results near 372 days ( $\log_{10} = 2.57$ ) due largely to image artifacts that repeat on the annual *Kepler* thermal cycle; this was also present in the Q1–Q16 results (Tenenbaum et al., 2014). In both Q1–Q17 runs there are peaks attributable to bright, short period sources in the *Kepler* field of view. The peaks near 0.57 days ( $\log_{10} = -0.24$ ) and 12.45 days ( $\log_{10} = 1.10$ ) are due to contamination by RR Lyrae and V380 Cyg respectively (Coughlin et al., 2014).

Figure 9.14 shows the TPS MES versus orbital period in days for the DR25 TCEs. The axes are displayed on logarithmic scales. All 34,032 TCEs are shown in the top panel. The bottom panel displays the density of the distribution. There is a dense band of TCEs associated with relatively low detection statistics across the full range of periods. There is an excess of long-period, low MES TCEs dominated by false alarms. As discussed earlier, the vetoes are not as effective at long orbital periods with relatively few transit events. Furthermore, false alarm probability increases for long orbital periods because noise statistics based on relatively few transit events are not approximated well by a Gaussian distribution and we did not increase the transit detection threshold to maintain a uniform false alarm rate (as was effectively enforced with the bootstrap veto in DR24).

Figure 9.15 shows the distribution of DR25 MES displayed on a linear scale. The high end of the MES distribution is clipped. In the left panel we see 31,064 TCEs with MES below  $10\sigma$ ; in the right panel we see 27,251 TCEs with MES below  $20\sigma$ . The mode of the distribution is at  $8\sigma$  whereas Seader et al. (2015) reported a mode near  $9\sigma$  in the Q1–Q17 DR24 results. Modifications to the TPS search algorithm and consistency tests have produced a population of TCEs for which the mode is  $1\sigma$  closer to the transit detection threshold.

Figure 9.16 shows a histogram of transit duty cycles for the DR25 TCEs. The transit duty cycle is defined to be the ratio of the trial transit pulse duration to the detected period of the TCE (effectively the fraction of time during which the TCE is in transit). The TPS employs a configurable upper limit on the duty cycles searched. The limit for this search was 0.16. This implies that the search for short-period transit signals does not include all of the trial pulse durations discussed earlier. The large number of short-period TCEs produces a ramp in duty cycle from 0.05 to 0.16 as described by Seader et al. (2015). The large number of long-period TCEs in the DR25 results, however, corresponds to many more TCEs at the lowest duty cycles than were generated in the DR24 run.



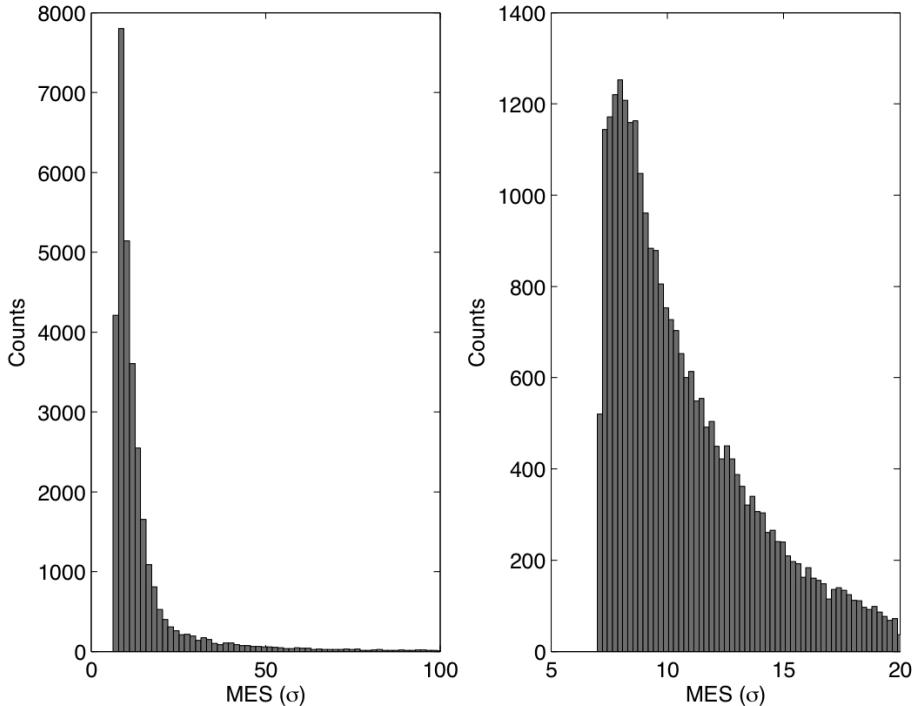
**Figure 9.14** Multiple Event Statistic (MES) versus orbital period in days on logarithmic scales. Top: All DR25 TCEs. Bottom: Density for all TCEs on logarithmic scale. From Figure 6 of Twicken et al. (2016).

#### 9.5.4 Comparison with Known Kepler Objects of Interest (KOIs)

The performance of the latest Q1–Q17 pipeline run may be evaluated based on the rate of recovery of a set of well established, high quality KOIs. As in the past, we refer to these as “golden KOIs.” The ephemerides and dispositions for the golden KOIs were obtained from the cumulative KOI table at the NASA Exoplanet Archive on 2015 September 25, after the DR24 KOI table was finalized. The cumulative KOI table has been aggregated from past transit searches and *Kepler* planet catalogs published by Borucki et al. (2011a,b), Batalha et al. (2013), Burke et al. (2014), Rowe et al. (2015), Mullally et al. (2015), and Coughlin et al. (2016). Selection criteria for the golden KOIs were:

1. MES above  $9.0\ \sigma$  in the most recent *Kepler* pipeline transit search in which the KOI was recovered, and
2. Disposition as Planet Candidate (PC) following two or more prior transit searches including at least one of Q1–Q16 and Q1–Q17 DR24. The DR25 golden KOI set includes 3402 KOIs on 2,621 unique target stars. The size of this set far exceeds the number of golden KOIs employed in the past for evaluating pipeline TPS performance. There were 1752 golden KOIs on 1483 target stars in the Q1–Q17 DR24 analysis, so the number of test cases for evaluating transit search performance nearly doubled for the final Q1–Q17 run.

TPS does not receive prior knowledge regarding KOIs or detections on specific targets. Recovery of objects of interest that were previously detected is a valuable test to guard against inadvertent introduction of significant flaws into the detection algorithm. The DR25 golden KOI set was selected for evaluation of the SOC 9.3 codebase and associated pipeline runs approximately six months prior to the TPS and DV runs that produced the TCE population discussed here. Of the 3402 golden KOIs, 3,385 (99.5%) were dispositioned in the cumulative KOI table as PC at the time of the DR25 transit search. These include a number of PCs in systems featuring



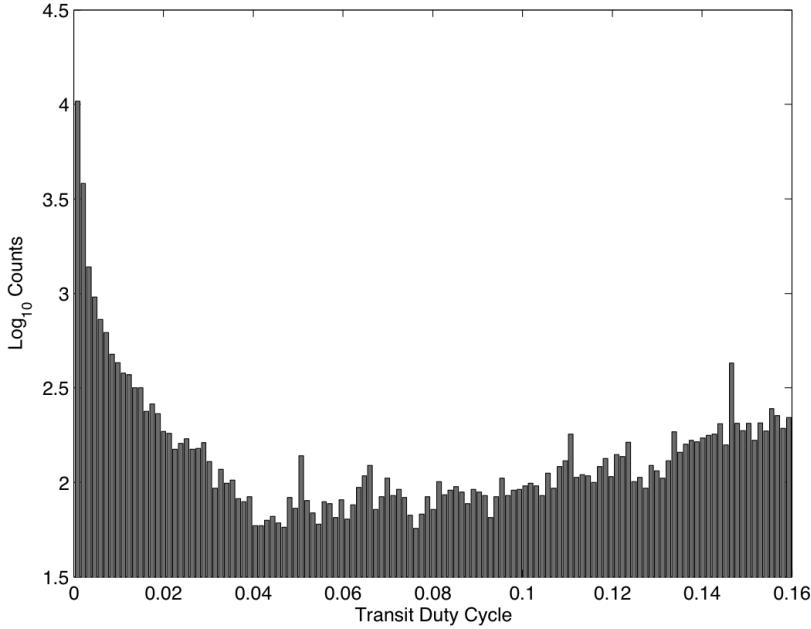
**Figure 9.15** Distribution of MES for all DR25 TCEs. Left: TCEs with MES below  $100\sigma$ . Right: TCEs with MES below  $20\sigma$ . From Figure 7 of Twicken et al. (2016).

Transit Timing Variations (TTV). The pipeline was not designed to detect transit signatures with TTVs and has never been upgraded specifically for that purpose. Detection of transiting planets with TTVs represents a test of the robustness of the search algorithm, which assumes strictly periodic transit signals.

Although TPS does not receive prior knowledge regarding KOIs, DV is provided with the ephemerides (period, epoch, and transit duration) of the individual KOIs associated with each of the targets that produce TCEs in the transiting planet search. KOI ephemerides are only employed by DV to match pipeline results for individual TCEs to known KOIs as an aid to *Kepler* project personnel and the greater science community. Matches to known KOIs (at the time that DV is run) are displayed in DV Reports by target and DV Report Summaries by TCE; these pipeline products are delivered to the NASA Exoplanet Archive. KOI ephemerides are not employed by DV for any purpose other than matching TCEs produced in the current pipeline run to previously known KOIs. An alternative matching algorithm is utilized outside of the pipeline to federate new TCEs with known KOIs when creating KOI tables at the Exoplanet Archive. It is possible, and even likely, that there will be discrepancies between KOI matching results displayed in DR25 DV Reports and DV Report Summaries and the DR25 KOI table at the Archive. The matching of TCEs to KOIs is discussed further in Subsection 9.5.5.

It should be noted that PDC does employ ephemerides of known KOIs and eclipsing binaries in conditioning data for the TPS. The ephemerides are used to identify in-transit (or in-eclipse) cadences for given targets; data samples for those targets and cadences are subsequently protected from misidentification as outliers or Sudden Pixel Sensitivity Dropouts (SPSD) (Jenkins et al., 2010a; Stumpe et al., 2012).

Figure 9.17 shows the distribution of transit depth, SNR, period and planet radius for the DR25 golden KOIs. Parameters are displayed on logarithmic scales. As discussed earlier, the



**Figure 9.16** Distribution of transit duty cycles on logarithmic scale for the 34,032 DR25 TCEs. Duty cycle is defined as the ratio of the trial transit pulse duration to the detected period of the TCE. From Figure 8 of Twicken et al. (2016).

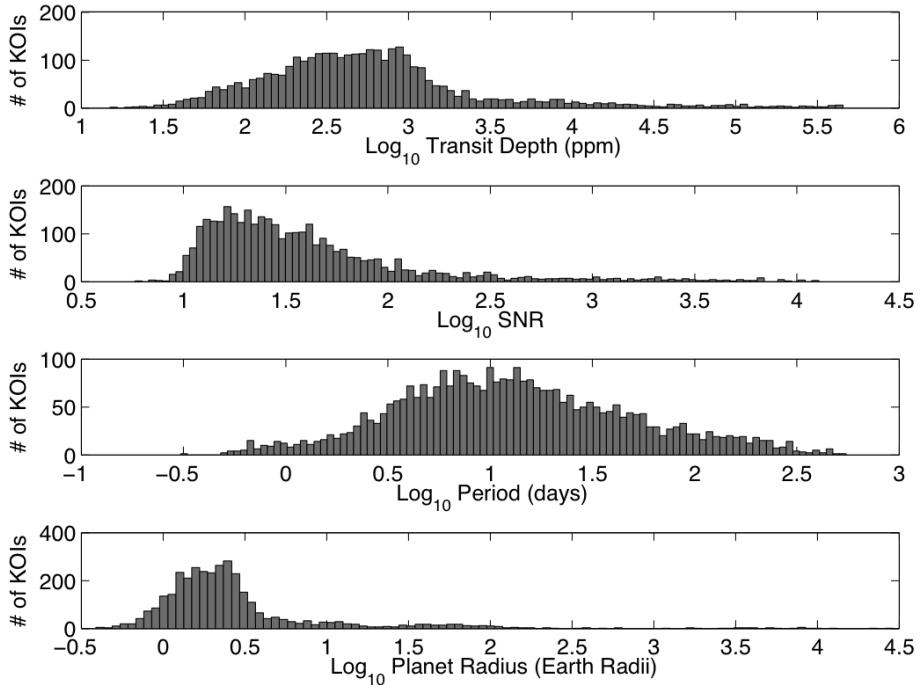
parameter values were obtained from the cumulative KOI table at the NASA Exoplanet Archive on 2015 September 25. The golden KOIs span a large region of transiting planet parameter space. In the final pipeline run, TPS produced TCEs on targets hosting 3397 of the 3402 golden KOIs. TCEs were not produced on targets hosting the following golden KOI:

- 4253.01 (period 173.3d, SNR 13.5)
- 4670.01 (period 6.81d, SNR 11.5)
- 4886.01 (period 18.0d, SNR 12.1)
- 5727.01 (period 65.4d, SNR 9.6)
- 5850.01 (period 303.2d, SNR 12.7).

The TPS search latched onto the correct period for all five of these KOIs. The reasons for failure to produce TCEs on these targets include:

- Maximum MES detection statistic ( $7.04\sigma$ ) was below the  $7.1\sigma$  threshold (KOI 5727.01).
- Veto by  $\chi^2_{(GOF)}$  consistency test (4670.01, 4886.01, and 5850.01).
- Veto by both  $\chi^2_{(2)}$  and  $\chi^2_{(GOF)}$  consistency tests (4253.01).

The  $\chi^2$  vetoes were discussed earlier and described in detail by Seader et al. (2013). We will now discuss matching of KOI and TCE ephemerides for the targets hosting golden KOIs on which TCEs were generated in the DR25 run.



**Figure 9.17** Parameter distributions of “golden KOIs” employed to assess performance of DR25 transit search. Parameters are displayed on logarithmic horizontal axes. Parameter values were obtained from the cumulative KOI table at the NASA Exoplanet Archive on 2015 September 25. From Figure 9 of Twicken et al. (2016).

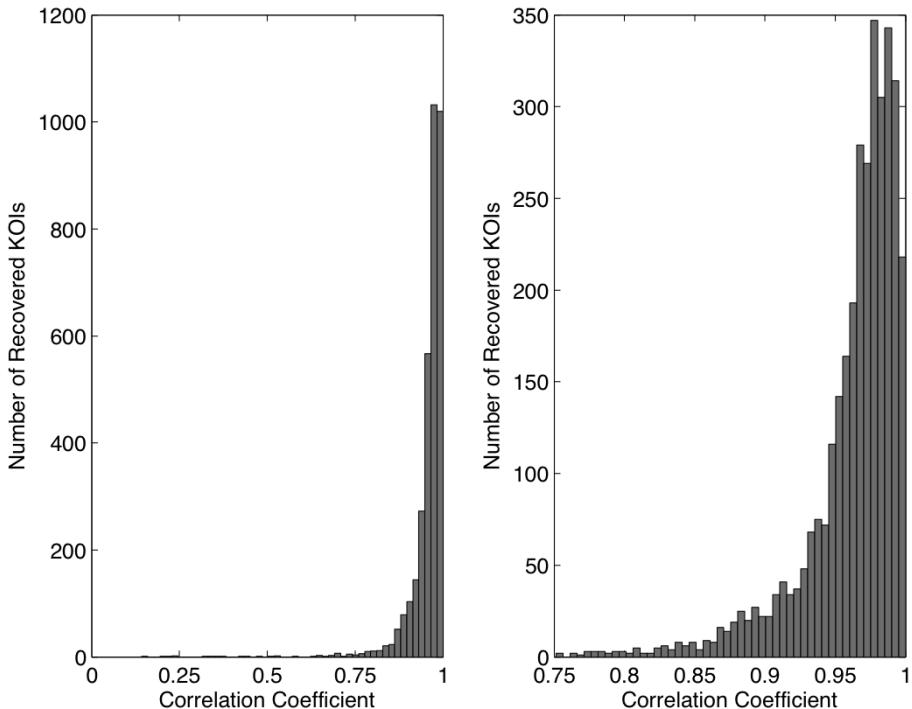
### 9.5.5 Matching of Golden KOI and TCE Ephemerides

Generation of a TCE on a target hosting a golden KOI is not sufficient to determine that the KOI has been recovered. It is necessary that the transit ephemeris produced in the pipeline match that of the KOI, such that the transit events producing the pipeline detection are consistent with the KOI ephemeris. As stated earlier, DV is provided with the ephemerides (period, epoch, transit duration) of each of the previously known KOIs associated with the targets that produce TCEs in the TPS. Once the limb-darkened model fitting and search for additional planets have concluded, DV matches each of the pipeline results against the ephemerides of the known KOIs. The matching is performed by correlating high-temporal resolution, rectangular KOI transit waveforms against rectangular transit waveforms based on pipeline ephemerides. Pearson correlation coefficients are compared against an ephemeris matching threshold (0.75 for 0 matching in DV) to establish whether or not a match has been produced. The transit waveforms are normalized such that correlation coefficients equal to 1.0 indicate perfect matches and correlation coefficients equal to 0.0 indicate that KOI and pipeline transit events are non-overlapping. The matching threshold was specified at a level that is likely to be reached only if a KOI is accurately recovered. A match is not declared in DV if the correlations exceed the matching threshold between a single TCE and multiple KOIs on a given target (which does happen in the case of duplicate KOIs) or if the correlations exceed the matching threshold between multiple TCEs on a given target and a single KOI.

In the DR25 run, DV reported an ephemeris match at the specified threshold or better for 3354 of 3402 golden KOIs. These golden KOIs may be assumed to have been recovered without further investigation. Of the 48 golden KOIs that did not trigger an ephemeris match, we have

seen that in 5 cases there were no TCEs on the host target. We investigated the remaining 43 golden KOIs for which a TCE was produced on the host target but a match was not reported by DV in order to ascertain whether the KOI in question had, in fact, been recovered. We found that 40 of the 43 golden KOIs in this category were indeed recovered in the DR25 run; details are provided below. In total, 3394 of 3402 DR25 golden KOIs (99.8%) were recovered in the pipeline with the SOC 9.3 codebase.

Figure 9.18 shows the distribution of ephemeris match correlation coefficients for the 3394 golden KOIs that were recovered in the DR25 pipeline run. The full range of correlation coefficients is displayed in the left panel and the range of correlation coefficients above the pipeline ephemeris matching threshold (0.75) is displayed in the right hand panel. Most of the ephemerides were matched at a high level. In fact, 92.0% of the golden KOIs that were recovered in the run produced correlation coefficients  $> 0.9$ .



**Figure 9.18** Distribution of ephemeris match correlation coefficients for “golden KOIs” recovered in DR25 run. Left: All recovered KOIs. Right: Recovered KOIs at ephemeris matching threshold (0.75) and above. From Figure 10 of Twicken et al. (2016).

Table 9.1 lists the complete golden KOI set, KOI and TCE ephemerides, and corresponding ephemeris match correlation coefficients; the table is sorted by KOI number. The time standard for epoch of first transit is Barycentric Kepler-modified Julian Date (BKJD). It should be noted that DV employs a standard convention for reporting epochs (first transit after start of Q1) whereas the cumulative KOI table at the Exoplanet Archive does not; hence, there are often an integer number of orbital periods between the KOI and TCE epochs. TCE ephemerides and correlation coefficients are listed as -1 for the eight golden KOIs that were not recovered in the DR25 run.

The five golden KOIs for which no TCE was produced on the host target were discussed earlier. The three unrecovered golden KOIs for which at least one TCE was produced on the host target are:

**Table 9.1** Golden KOI and TCE Ephemeris Matching Results

KOI Number	KOI Period (days)	TCE Period (days)	KOI Epoch (BKJD)	TCE Epoch (BKJD)	KOI Duration (hours)	TCE Duration (hours)	Correlation Coefficient
1.01	2.4706	2.4706	122.7633	132.6457	1.7426	1.7968	0.985
2.01	2.2047	2.2047	121.3586	132.3833	3.8822	4.0438	0.980
3.01	4.8878	4.8878	124.8131	134.5888	2.3639	2.4059	0.991
4.01	3.8494	3.8494	157.5267	134.4299	2.6605	2.7133	0.990
5.01	4.7803	4.7803	132.9741	132.9740	2.0349	2.0611	0.994
7.01	3.2137	3.2137	123.6119	133.2543	3.9935	4.1268	0.984
10.01	3.5225	3.5225	121.1194	131.6870	3.1906	3.2711	0.988
12.01	17.8552	17.8552	146.5964	146.5961	7.4294	7.4400	0.999
13.01	1.7636	1.7636	120.5659	132.9115	3.1814	3.1796	0.998
17.01	3.2347	3.2347	121.4866	134.4253	3.6011	3.5910	0.999
18.01	3.5485	3.5485	122.9015	133.5470	4.5770	4.5615	0.998
20.01	4.4380	4.4380	171.0091	135.5055	4.7010	4.6918	0.999
22.01	7.8914	7.8914	177.2500	137.7928	4.3040	4.3537	0.994
41.01	12.8159	12.8159	122.9482	135.7632	6.3728	6.5001	0.990
41.02	6.8871	6.8871	133.1779	133.1786	4.4276	4.5472	0.987
41.03	35.3331	35.3331	153.9833	153.9855	5.9040	6.0418	0.982
42.01	17.8337	17.8338	181.2337	145.5634	4.5403	4.8219	0.970
46.01	3.4877	3.4877	170.9320	132.5683	3.8427	3.9487	0.987
46.02	6.0298	6.0297	132.4818	132.4809	3.9100	4.0465	0.952
49.01	8.3138	8.3138	175.9916	134.4234	2.9927	3.1174	0.980

Table 9.1 is published in its entirety in a machine readable format in Twicken et al. (2016). A portion is shown here for guidance regarding its form and content.

- 989.03 (period 16.2d, SNR 52.8),
- 2048.02 (period 99.7, SNR 12.6), and
- 3051.01 (period 11.7d, SNR 11.2).

The reasons for failure to produce TCEs for these KOIs include:

- Self-timeout in DV model fitting process on earlier TCEs associated with false positive KOIs on target (KOI 989.03).
- Veto by both  $\chi^2_{(2)}$  and  $\chi^2_{(GOF)}$  consistency tests after latching onto correct period (2048.02 and 3051.01).

We explain below why a match was not reported at the specified threshold in DV for many of the 40 golden KOIs that were recovered to provide a sense of the difficulties involved in benchmarking pipeline completeness by matching TCEs to KOI ephemerides:

- KOI and DV periods differ by integer factor and KOI period appears to be incorrect (2174.03, 4829.01, 4893.01).
- KOI and DV periods differ by integer factor and DV period appears to be incorrect (2306.01, 2732.04).

- KOI and DV periods differ by integer factor but true period is ambiguous because target was observed only in every other quarter (5568.01).
- KOI period is less than 0.5 days and DV produces two TCEs at twice the true period because the minimum search period is 0.5 days (2916.01).
- KOI transit duration is less than one cadence; minimum TPS trial transit duration is 1.5 hours (three cadences) and DV fitter declares error if duration derived from fit parameters is less than one cadence (4546.01).
- KOI is in a TTV system (Ford et al., 2012; Mazeh et al., 2013) where linear KOIs and pipeline ephemerides only approximate observed transit signal (KOI 277.02, 456.02, 884.02, 984.01, 1831.03).
- KOI was observed late in mission only; discrepancy exists between KOI and DV epochs when projected back to start of science operations (5403.01, 5605.01, 5672.01, 6145.02, 6166.02).
- Duplicate KOIs exist; DV does not report match because pipeline result actually matches two different KOIs on same target (1101.01, 2768.01).
- KOI is binary (Kirk et al., 2016) featuring deep eclipses (>25%) that DV does not fit by design; TPS ephemeris does not match KOI at specified threshold (3545.01, 3554.01, 5797.01).
- KOI appears to be heartbeat star (Kirk et al., 2016) that does not feature conventional transits or eclipses but rather rings due to tidal pulsations (2215.01).

In summary, 3394 of 3402 golden KOIs (99.8%) selected in advance for assessment of the performance of the DR25 run with the SOC 9.3 codebase were recovered in the pipeline. Of the eight golden KOIs that were not recovered, in one case the correct period was latched but there was a failure to produce a detection statistic at the required threshold (7.04 versus  $7.1\sigma$ ), in six cases the correct period was latched and a sufficient detection statistic was produced but a TCE was vetoed by one or two of the  $\chi^2$  transit consistency tests, and in one case a timeout limit was reached while processing prior false positive TCEs in a multiple-KOI system.

The vetoes are a necessary evil for the TPS. They impact completeness to some degree. Without the vetoes, however, the sheer number of TCEs would overwhelm the ability to process them in the pipeline and to produce a reliable catalog of planetary candidates. As stated earlier, the DR25 run would have produced TCEs on 68,004 unique targets at the  $7.1\sigma$  level or greater in the absence of the vetoes. This would have approximately doubled to 136,000 total TCEs after the search for additional planets. A larger number of TCEs does not necessarily imply better completeness, however; the vetoes allow continued search of light curves without passage to DV and removal of flux data in the pipeline search for multiple planets on individual targets. It is interesting to note that there were 126,153 unique targets with MES above  $7.1\sigma$  prior to application of the vetoes in the DR24 run (Seader et al., 2015); this is significantly larger than the 68,004 targets with MES above  $7.1\sigma$  in the DR25 run. We believe that the reduction in DR25 is likely attributable to implementation of quarter-by-quarter whitening in TPS discussed earlier.

The 99.8% recovery rate with the SOC 9.3 codebase represents exceptional performance for a large set of KOIs. Seader et al. (2015) reported that 1,664 of 1,752 golden KOIs (95.0%) were recovered in the Q1–Q17 DR24 run with the SOC 9.2 codebase. Tenenbaum et al. (2014) reported that 1,597 of 1,646 golden KOIs (97.0%) were recovered in the Q1–Q16 run with the SOC 9.1 codebase.

Although recovery of a large set of established KOIs is a solid test of the performance of the Pipeline TPS, it should be noted that KOIs do not reflect ground truth as the true nature of these

objects of interest is not actually known. A better measure of the detection efficiency of the pipeline is the recovery of transiting planet signatures injected into *Kepler* flight data for a large number of targets over a range of orbital periods and planet radii (and hence SNR). Such studies have been performed for 4 quarters (Q9–Q12) with the SOC 9.1 codebase (Christiansen et al., 2015) and 17 quarters with the SOC 9.2 codebase (Christiansen et al., 2016). These studies involve injection of transit signatures into calibrated pixel data and subsequently running the *Kepler* Pipeline through the PA, PDC, TPS, and DV components. Pixel-level injections also offer the opportunity to characterize the performance of the Data Validation diagnostics employed to help differentiate between true transiting planet signatures and false positive detections attributable to eclipsing binaries and background sources (Bryson et al., 2013; Mullally et al., 2015; Coughlin et al., 2016). A pixel-level transit injection run is currently underway to assess the detection efficiency of the DR25 pipeline (Christiansen 2016, in Prep.).

Parallel studies are also underway involving the injection of transit signals into the systematic error-corrected light curves of a representative sample of LC targets. The flux-level injections are repeated many times (typically  $> 600,000$ ) for each target in order to deeply probe TPS detection efficiency over a range of orbital periods and planet radii. Both pixel-level and flux-level injections represent a superior gauge of pipeline completeness compared to recovery of existing KOIs because these are controlled experiments where a very large number of transit signatures may be injected and ground truth is available. Nevertheless, transit injections require considerable time and computational resources; assessing pipeline performance based on recovery of established KOIs remains a valuable exercise and is an efficient way to compare and contrast performance from one transit search to the next.

## 9.6 Conclusions

The Transiting Planet Search pipeline module has seen significant development and improvements over the six years it has been in operation since *Kepler*'s launch in March 2009. Non-ideal behavior of the instrument and complex stellar photometric variability motivated these changes. SOC 9.3 is the most sensitive version of TPS to date and has delivered over 34,000 transit-like features to the NExScI archive.

## Appendix A: Generation of the Non-Decimated Octave-Band Filters

This appendix defines how the bandpass filters used by TPS are generated. In the more typical case for wavelet analysis, the implemented wavelet transform is critically sampled, meaning that if the input time series has  $N$  points, then the output wavelet transform vector is  $N$  points as well. In this case, the high-pass filter,  $h_1(n)$ , and the low-pass filter,  $h_0(n)$ , retain the same form throughout the transformation process and the low pass-filtered signal,  $x_j(n)$ , is decimated at each stage prior to separating the resulting time series into a “high-pass” and a “low-pass” time series.

Figure 9-A.1 shows the first two stages of a classical, critically sampled wavelet expansion of a time series,  $x(n)$ . The original time series,  $x(n)$ , is fed through a high-pass filter with frequency response,  $H_1(\omega)$ , called the mother wavelet, then down sampled by a factor of 2 to yield the highest bandpass-filtered output time series,  $x_1(n)$ . The original time series is also subjected to a low-pass filter with a frequency response,  $H_0(\omega)$ , called the father wavelet, then down sampled by a factor 2 before entering the second stage of the transform. The process continues until the output time series is no longer than the bandpass filters. In this classical wavelet transform, each output time series,  $x_j(n)$ , is half the length of the previous one, and the total number of points

in the set of output time series is the same as the number of points in the input time series. The filters in the filter bank representation in Figure 9.4 can be read off from Figure 9-A.1 by scanning the sequence of filters and down-sample operations from  $x(n)$  to each output time series. For example,  $H_1$  of Figure 9.4 corresponds to the mother wavelet  $H_1$  followed by the down-sample by 2 operation. Likewise,  $H_2$  corresponds to the sequence  $H_0$ , followed by down-sampling by 2, followed by  $H_1$ , followed by down-sampling by 2.

For the TPS application, however, preserving the property of shift invariance is much more important than avoiding the increase in memory required to hold the results of a non-decimated wavelet transform. Deriving the filters for the over-complete wavelet transform requires that we “pull” the filters in Figure 9-A.1 through the preceding “down-sample by 2” operations so that the result of each stage is non-decimated and therefore the same length as the input time series.

The mother wavelet,  $h_1$ , can be derived from  $h_0$  by time reversal and alternating the sign of each tap:

$$h_1(n) = (-1)^n h_0(-n). \quad (\text{A.1})$$

It is convenient to represent the resulting filters in the frequency domain. Let  $H_0^j$  and  $H_1^j$  be the  $N$ -point Discrete Fourier Transforms of  $h_0^j(n)$  and  $h_1^j(n)$ , respectively for stage  $j$ . It can be shown that the band pass filters at the next stage are related to the current stage by decimating the frequency response of each filter and concatenating it with itself to keep the same number of points in the vector:

$$H_0^{j+1}(n) = \begin{cases} H_0^j(2n), & n = \{0, \dots, (N-1)/2\} \\ H_0^j(2n-N+1), & n = \{N/2, \dots, N\}, \end{cases} \quad (\text{A.2})$$

and

$$H_1^{j+1}(n) = \begin{cases} H_1^j(2n), & n = \{0, \dots, (N-1)/2\} \\ H_1^j(2n-N+1), & n = \{N/2, \dots, N\}. \end{cases} \quad (\text{A.3})$$

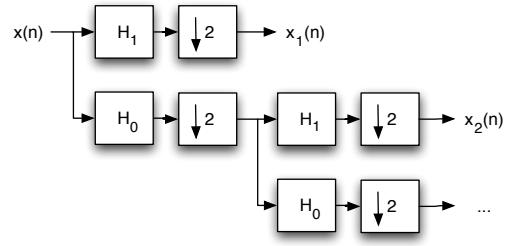
In MATLAB pseudocode, this would be

$$\begin{aligned} H'_0 &= [H_0(1 : 2 : end); H_0(1 : 2 : end)]; \\ H'_1 &= [H_1(1 : 2 : end); H_1(1 : 2 : end)]; \end{aligned} \quad (\text{A.4})$$

The first four filters in Figure 9.4 are then given by:

$$\begin{aligned} H_1(\omega) &= H_1^1(\omega) \\ H_2(\omega) &= H_0^1(\omega)H_1^2(\omega) \\ H_3(\omega) &= H_0^1(\omega)H_0^2(\omega)H_1^3(\omega) \\ H_4(\omega) &= H_0^1(\omega)H_0^2(\omega)H_0^4(\omega)H_1^3(\omega). \end{aligned} \quad (\text{A.5})$$

TPS constructs the bandpass filters “in place”, however, which is computationally convenient.



**Figure 9-A.1** Data flow diagram for a critically sampled wavelet expansion of a signal. The first two stages of the transform are shown here.

## Appendix B: Determining the Threshold for Positive Outlier Removal

In Subsection 9.3.2, an algorithm is described that removes positive outliers from each flux time series prior to searching same for transits. The algorithm requires a threshold for removal of the outliers, set at  $12.3\sigma$ . The rationale behind this value is explained below.

The most obvious requirement for the positive outlier threshold is that it should, indeed, address only outliers and not flux values that are merely above the mean value due to statistical fluctuations. For a whitened flux time series with 69,810 samples, in the limit of Gaussian statistics, we expect that, on average, 1 sample will lie  $4.2\sigma$  above the mean. Therefore, any threshold significantly higher than  $4.2\sigma$  can be expected to be harmless in terms of its effect on sample values, which are driven by statistical fluctuations alone.

Another consideration is that the purpose of removing positive outliers is actually to suppress the formation of transit-like features in the ring-down of the outliers, since the latter can lead to false-alarm detections. Given a detection threshold of  $7.1\sigma$  and a requirement of at least 3 transits for detection, a negative feature with a significance of  $7.1\sigma \times \sqrt{3} = 12.3\sigma$  can exceed the detection threshold when paired with two “transits” that each have a significance of zero. This implies that a transit-like feature with a significance of  $12.3\sigma$  has a strong probability of triggering a false alarm, and thus our threshold for removing positive outliers should be set such that any outlier likely to have a  $12.3\sigma$  transit-like feature in its ring-down is removed.

Since the ring-down of a positive outlier will always be smaller than the outlier itself, it follows that removing positive outliers with  $12.3\sigma$  significance is more than sufficient to prevent formation of negative ring-down features with  $12.3\sigma$  significance. Since  $12.3\sigma$  is much larger than  $4.2\sigma$ , it also follows that this threshold will be benign from the point of view of ignoring statistical fluctuations. Thus a threshold of  $12.3\sigma$  was adopted for positive outlier removal.

## Bibliography

- Akaike, H., 1974. “A New Look at the Statistical Model Identification,” IEEE Transactions on Automatic Control, 19, 716
- Allen, B., 2005. “ $\chi^2$  Time-Frequency Discriminator for Gravitational Wave Detection,” Phys. Rev. D, 71, 062001
- Baggio, L., Cerdonio, M., Ortolan, A., et al., 2000. “ $\chi^2$  Testing of Optimal Filters for Gravitational Wave Signals: An Experimental Implementation,” Phys. Rev. D, 61, 102001
- Batalha, N. M., Rowe, J. F., Bryson, S. T., et al., 2013. “Planetary Candidates Observed by Kepler. III. Analysis of the First 16 Months of Data,” ApJS, 204, 24
- Borucki, W. J., Koch, D. G., Basri, G., et al., 2011. “Characteristics of Kepler Planetary Candidates Based on the First Data Set,” ApJ, 728, 117
- , 2011. “Characteristics of Planetary Candidates Observed by Kepler. II. Analysis of the First Four Months of Data,” ApJ, 736, 19
- Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al., 2013. “Identification of Background False Positives from Kepler Data,” PASP, 125, 889
- Burke, C. J., Bryson, S. T., Mullally, F., et al., 2014. “Planetary Candidates Observed by Kepler IV: Planet Sample from Q1-Q8 (22 Months),” ApJS, 210, 19
- Chandrasekaran, H. 2004. “Short Data Gap Filling Algorithm Prototype,” Tech. Rep. KADN-26067, NASA KPO@Ames Design Note

- Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al., 2012. "The Derivation, Properties, and Value of Kepler's Combined Differential Photometric Precision," *PASP*, 124, 1279
- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al., 2013. "Measuring Transit Signal Recovery in the Kepler Pipeline. I. Individual Events," *ApJS*, 207, 35
- , 2015. "Measuring Transit Signal Recovery in the Kepler Pipeline II: Detection Efficiency as Calculated in One Year of Data," *ApJ*, 810, 95
- , 2016. "Measuring Transit Signal Recovery in the Kepler Pipeline. III. Completeness of the Q1-Q17 DR24 Planet Candidate Catalogue with Important Caveats for Occurrence Rate Calculations," *ApJ*, 828, 99
- Claret, A., & Bloemen, S., 2011. "Gravity and Limb-Darkening Coefficients for the *Kepler*, CoRoT, Spitzer, uvby, UBVRIJHK, and Sloan photometric systems," *Astronomy & Astrophysics*, 529, A75, available online at <http://dx.doi.org/10.1051/0004-6361/201116451>
- Coughlin, J. L., Thompson, S. E., Bryson, S. T., et al., 2014. "Contamination in the Kepler Field. Identification of 685 KOIs as False Positives via Ephemeris Matching Based on Q1-Q12 Data," *AJ*, 147, 119
- Coughlin, J. L., Mullally, F., Thompson, S. E., et al., 2016. "Planetary Candidates Observed by Kepler. VII. The First Fully Uniform Catalog Based on the Entire 48-month Data Set (Q1-Q17 DR24)," *ApJS*, 224, 12
- Daubechies, I., 1988. "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, 41, 909
- Ford, E. B., Ragozzine, D., Rowe, J. F., et al., 2012. "Transit Timing Observations from Kepler. V. Transit Timing Variation Candidates in the First Sixteen Months from Polynomial Models," *ApJ*, 756, 185
- Gilliland, R. L., Chaplin, W. J., Jenkins, J. M., Ramsey, L. W., & Smith, J. C., 2015. "Kepler Mission Stellar and Instrument Noise Properties Revisited," *AJ*, 150, 133
- Gilliland, R. L., Brown, T. M., Guhathakurta, P., et al., 2000. "A Lack of Planets in 47 Tucanae from a Hubble Space Telescope Search," *ApJL*, 545, L47
- Gilliland, R. L., Chaplin, W. J., Dunham, E. W., et al., 2011. "Kepler Mission Stellar and Instrument Noise Properties," *ApJS*, 197, 6
- Jenkins, J. M., 2002. "The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets," *ApJ*, 575, 493
- Jenkins, J. M., Caldwell, D. A., & Borucki, W. J., 2002. "Some Tests to Establish Confidence in Planets Discovered by Transit Photometry," *ApJ*, 564, 495
- Jenkins, J. M., Doyle, L. R., & Cullers, D. K., 1996. "A Matched Filter Method for Ground-Based Sub-Noise Detection of Terrestrial Extrasolar Planets in Eclipsing Binaries: Application to CM Draconis," *Icarus*, 119, 244
- Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al., 2010. "Overview of the Kepler Science Processing Pipeline," *ApJL*, 713, L87

- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010b. “Transiting Planet Search in the Kepler Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D
- Jenkins, J. M., Twicken, J. D., Batalha, N. M., et al., 2015. “Discovery and Validation of Kepler-452b: A 1.6  $R_{\oplus}$  Super Earth Exoplanet in the Habitable Zone of a G2 Star,” AJ, 150, 56
- Kay, S., 1999. “Adaptive Detection for Unknown Noise Power Spectral Densities,” IEEE Trans. Signal Processing, 47, 10
- Kirk, B., Conroy, K., Prša, A., et al., 2016. “Kepler Eclipsing Binary Stars. VII. The Catalog of Eclipsing Binaries Found in the Entire Kepler Data Set,” AJ, 151, 68
- Koch, D. G., Borucki, W. J., Basri, G., et al., 2010. “Kepler Mission Design, Realized Photometric Performance, and Early Science,” ApJL, 713, L79
- Li, J., Allen, C., Bryson, S. T., et al. 2010. “Photometer Performance Assessment in Kepler Science Data Processing,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77401T
- Mandel, K., & Agol, E., 2002. “Analytic Light Curves for Planetary Transit Searches,” ApJL, 580, L171
- Mazeh, T., Nachmani, G., Holczer, T., et al., 2013. “Transit Timing Observations from Kepler. VIII. Catalog of Transit Timing Measurements of the First Twelve Quarters,” ApJS, 208, 16
- McCauliff, S., Cote, M. T., Girouard, F. R., et al. 2010. “The Kepler DB: A Database Management System for Arrays, Sparse Arrays, and Binary Data,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400M
- Mullally, F., Coughlin, J. L., Thompson, S. E., et al., 2015. “Planetary Candidates Observed by Kepler. VI. Planet Sample from Q1–Q16 (47 Months),” ApJS, 217, 31
- Rowe, J. F., Coughlin, J. L., Antoci, V., et al., 2015. “Planetary Candidates Observed by Kepler. V. Planet Sample from Q1–Q12 (36 Months),” ApJS, 217, 16
- Seader, S., Tenenbaum, P., Jenkins, J. M., & Burke, C. J., 2013. “ $\chi^2$  Discriminators for Transiting Planet Detection in Kepler Data,” ApJS, 206, 25
- Seader, S., Jenkins, J. M., Tenenbaum, P., et al., 2015. “Detection of Potential Transit Signals in 17 Quarters of Kepler Mission Data,” ApJS, 217, 18
- Stumpe, M. C., Smith, J. C., Van Cleve, J. E., et al., 2012. “Kepler Presearch Data Conditioning I – Architecture and Algorithms for Error Correction in Kepler Light Curves,” PASP, 124, 985
- Tenenbaum, P., Christiansen, J. L., Jenkins, J. M., et al., 2012. “Detection of Potential Transit Signals in the First Three Quarters of Kepler Mission Data,” ApJS, 199, 24
- Tenenbaum, P., Jenkins, J. M., Seader, S., et al., 2013. “Detection of Potential Transit Signals in the First 12 Quarters of Kepler Mission Data,” ApJS, 206, 5
- , 2014. “Detection of Potential Transit Signals in 16 Quarters of Kepler Mission Data,” ApJS, 211, 6
- Thompson, S. E., Jenkins, J. M., Caldwell, D. A., et al. 2016. “Kepler Data Release 25 Notes,” Tech. Rep. KSCI-19065-001, NASA Ames Research Center Kepler Mission

- Twicken, J. D., Jenkins, J. M., Seader, S. E., et al., 2016. “Detection of Potential Transit Signals in 17 Quarters of Kepler Data: Results of the Final Kepler Mission Transiting Planet Search (DR25),” *AJ*, 152, 158
- Van Trees, H. L. 1968, *Detection, Estimation, and Modulation Theory, Part I* (Wiley), 19–155, 239–442
- Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010. “Data Validation in the Kepler Science Operations Center Pipeline,” in *Proc. SPIE*, Vol. 7740, 42W

## CHAPTER 10

---

# A COMPUTATIONALLY EFFICIENT STATISTICAL BOOTSTRAP TEST FOR TRANSITING PLANETS

JON M. JENKINS<sup>1</sup>, SHAWN SEADER<sup>2</sup>, AND CHRISTOPHER J. BURKE<sup>2</sup>

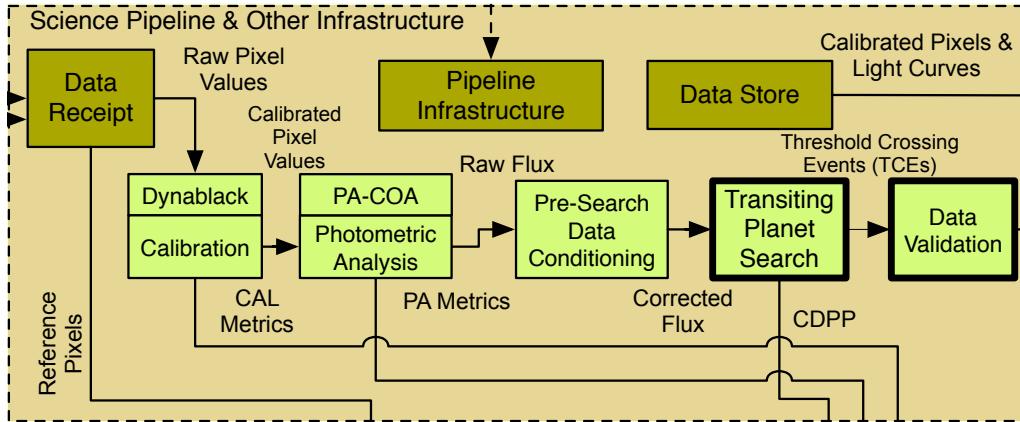
<sup>1</sup>NASA Ames Research Center, Moffett Field, CA 94035, <sup>2</sup>The SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** During its primary mission, the *Kepler* spacecraft nearly continuously observed a host of  $\sim 165,000$  target stars over a four-year period to detect transiting planets. Applying a  $7.1\sigma$  detection yields over 100,000 detections in each search, the vast majority of which are false alarms. Applying a set of false alarm vetoes described in Seader et al. (2015) eliminates a vast majority of the false alarms. For the remaining detections, a bootstrap procedure described in Jenkins (2002) was proposed to estimate the distribution of the null statistics and thereby estimate the false alarm probability to both further discriminate against false alarms as well as boost confidence in true detections. This bootstrap procedure becomes computationally intractable, however, when the size of the dataset is large or the orbital period of the detected event is sufficiently small. This paper describes a new bootstrap procedure that permits efficient construction of the distribution of null statistics for the full range of detection scenarios. While this test was not applied as a veto in Transiting Planet Search (TPS), it is furnished as a diagnostic for all Threshold Crossing Events (TCEs) and was computed for the SOC 9.1 Q1–Q16, the Q1–Q17 DR24, and the SOC 9.3 Q1–Q17 DR25 transit searches. The bootstrap results are archived at NASA’s Exoplanet Science Institute’s (NExSci) exoplanet archive.

### 10.1 Introduction

The *Kepler* spacecraft continuously observed more than  $\sim 165,000$  target stars in a 116 square-degree field of view to discover Earth-like planets transiting Sun-like stars via analysis of photometric data (Borucki et al., 2010; Koch et al., 2010). The spacecraft collected photometric data for each target star, which was compressed and stored onboard to be downlinked at monthly intervals. The *Kepler* Science Operations Center (SOC) at NASA Ames Research Center processes the data with the Science Processing Pipeline, which is composed of several modules, including the Transiting Planet Search (TPS) (Jenkins et al., 2010), as illustrated in Section 10.1. To search for transit signatures, TPS employs a bank of wavelet-based matched filters that form a grid on a three-dimensional parameter space of transit duration, period, and epoch (Jenkins, 2002; Jenkins et al., 2010). A detection statistic, referred to as the Multiple Event Statistic (MES), is calculated for each template and compared to a threshold value  $\eta = 7.1\sigma$ . Owing to non-stationary and non-Gaussian noise, uncorrected systematics, and poorly mitigated noise events of either astrophysical or non-astrophysical nature, spurious Threshold Crossing Events (TCEs) can be produced by the matched filtering performed in TPS. The vetoes described in Seader et al. (2013) efficiently remove a majority of the false alarms associated with noise glitches. This

leaves a different class of false alarms that come about more generally by a failure of the noise process to meet the underlying assumptions made in the detection theory.



**Figure 10.1** The bootstrap algorithm is called as a diagnostic test in the Data Validation (DV) component of the *Kepler* Science Processing Pipeline. It provides a statistical model of Transiting Planet Search (TPS) for the purpose of estimating the statistical confidence in each Threshold Crossing Event (TCE) detected by TPS in both the original run and in the multiple planet search iterations within DV. The bootstrap analysis is conducted on the residual light curve obtained at the end of the DV process by masking out the transits of each TCE.

Detections in TPS are made under the assumption that the pre-whitening filter applied to the light curve yields a time series whose underlying noise process is stationary, white, Gaussian, and uncorrelated. When the noise deviates from these assumptions, the detection thresholds are invalid and the false alarm probability associated with such a detection may be significantly higher than that for a signal embedded in white Gaussian noise. So, not only is it important to understand the false alarm probability for a given detection but it is important to also understand whether the distribution of null statistics deviates from its assumed form. The Statistical Bootstrap Test, or the Bootstrap, is a way of building the distribution of the null statistics from the data so that the false alarm probability can be calculated for each TCE based on the distribution of the out-of-transit statistics. Then, given the probability distribution, one can also calculate the threshold needed to achieve a false alarm probability equivalent to that corresponding to using the search threshold on the assumed standard normal distribution (i.e.  $\sim 6.24 \times 10^{-13}$  for a threshold of  $7.1\sigma$  for a standard normal distribution). Comparing the detection statistic to this Gaussian noise threshold is an empirical way to relax some of the assumptions on the effectiveness of the whitener and assess the reliability of the detection.

Jenkins et al. (2002) formulated a bootstrap test for establishing the confidence level in planetary transit signatures identified in transit photometry in white, but possibly non-Gaussian, noise. Jenkins (2002) extended this approach to the case of non-white noise. In both cases, the bootstrap false alarm rate (FAR) as a function of the MES of the detected transit signature was estimated by explicitly generating individual bootstrap statistics directly from the set of out-of-transit data. This direct bootstrap sampling approach can become extremely computationally intensive as the number of transits for a given TCE grows beyond  $\sim 15$ . The number of individual bootstrap statistics that can be formed from the  $m$  out-of-transit cadences of a light curve and the  $p$  transits is  $m^p$ , which is  $\sim 2.9 \times 10^{48}$  statistics for 10 transits and 4 years of *Kepler* data. An alternative, more computationally efficient method can be implemented by formulating the bootstrap distribution in terms of the probability density function (PDF) of the single event detection statistics. The distribution for the MES as a function of threshold and number of transits can then be obtained from the distribution of single event statistics treated as a bivariate random process.

This paper is organized as follows. Section 10.2 provides an overview of the detection theory employed by TPS. Section 10.3 describes in detail a new bootstrap algorithm for estimating the distribution of the null statistics. Section 10.4 lists and describes the bootstrap algorithm products on the NExSci archive. Section 10.5 discusses and compares the results of the bootstrap algorithm as applied to three datasets: the SOC 9.1 (Q1–Q16) transit search (Tenenbaum et al., 2014, Subsection 10.5.1),<sup>1</sup> the SOC 9.2 (Q1–Q17) transit search (Seader et al., 2015, Subsection 10.5.2),<sup>2</sup> and the SOC 9.3 (Q1–Q17 DR25) transit search (Twicken et al., 2016, Subsection 10.5.3).<sup>3</sup> Section 10.6 describes a set of Monte Carlo experiments whose purpose is to examine the statistical precision of the bootstrap algorithm by analysis of white Gaussian noise (WGN) light curves that have been processed through TPS. Section 10.7 provides a graphical example of the bootstrap algorithm applied to a single light curve. Finally, Section 10.8 summarizes the main results.

## 10.2 Detection Algorithm

The data input into the search in TPS are discrete, contiguous, flux fraction time series that have been corrected for systematics, have had some level of harmonics removed, and have had some other more localized noise artifacts removed, such as sudden pixel sensitivity dropouts (SPSD), cosmic rays, or thermal transients. For a discussion of how the data are prepared for the search in TPS (see e.g., Chapter 9; Tenenbaum et al. (2012); Twicken et al. (2016)). Since we are observing the light from many target stars with highly varied properties, the noise  $w(n)$  is typically not white, but colored. To whiten the data, a joint time-frequency wavelet decomposition is performed on the data so that the noise can be estimated for each of several time-frequency bands using a moving circular median absolute deviation. For efficiency, the actual detection algorithm employed by TPS is a wavelet-based matched filter that is calculated in this wavelet domain, employing the noise estimates to whiten both the data and the templates. In the subsequent formulation of our detection algorithm, however, it is assumed that the noise is uncorrelated WGN with zero mean and unit variance:

$$\langle \tilde{w}(n) \rangle = 0 \text{ and} \quad (10.1)$$

$$\langle \tilde{w}(n) \tilde{w}(m) \rangle = \delta(n - m), \quad (10.2)$$

where  $\delta(n)$  is the Dirac delta function. So we assume that the data and templates have been whitened already and work in the more convenient time domain. The two formulations are equivalent. In reality, however, the whitening process is not perfect so the noise never strictly conforms to our assumptions (which is why the bootstrap is useful). Note that although individual transits are localized prior to the whitening, after doing the wavelet decomposition, whitening, and inverse wavelet decomposition to get back to the time domain, the transits have a larger extent. The contribution from out-of-transit data points to the MES for any given transit can be significant, so the full dot products in what follows must be carried out. For details of the whitening process and the construction of the wavelet-based matched filter output, see Chapter 9; Seader et al. (2015); Jenkins (2002).

<sup>1</sup>[http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1\\_q16\\_tce](http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1_q16_tce)

<sup>2</sup>[http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1\\_q17\\_dr24\\_tce](http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1_q17_dr24_tce)

<sup>3</sup>[http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1\\_q17\\_dr25\\_tce](http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=q1_q17_dr25_tce)

Let  $x(n)$  be this discrete, contiguous, whitened, flux fraction time series, where  $n \in [1, \dots, N]$ . Under the null hypothesis,  $H_0$ , there is no transit signal present and we have only noise  $w(n)$ . Under the alternative hypothesis,  $H_1$ , there is a transit pulse signal  $s(n)$  present in the data (for simplicity, assume for now a single pulse is present rather than a pulse train). We then have:

$$\begin{aligned} H_0 : \mathbf{x} &= \mathbf{w} \\ H_1 : \mathbf{x} &= \mathbf{w} + \mathbf{s}, \end{aligned} \quad (10.3)$$

where bold face indicates a vector quantity,  $\mathbf{a} = \langle a(1), a(2), \dots, a(N) \rangle$ . Given the assumed properties of the noise, the PDF of the noise under the two hypotheses are:

$$p_0(\mathbf{w}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mathbf{x} \cdot \mathbf{x}\right) \text{ and} \quad (10.4)$$

$$p_1(\mathbf{w}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{s}) \cdot (\mathbf{x} - \mathbf{s})\right), \quad (10.5)$$

where the dot product of two time series vectors is given by:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^N a(i)b(i). \quad (10.6)$$

In this detection scenario, there is no knowledge of prior probabilities nor any costs associated with false alarms or false dismissals, so the strategy we adopt is to use the Neyman-Pearson criteria. Using this criteria, the likelihood ratio  $\Lambda$  is calculated and compared with some threshold  $\Lambda_0$ . The value of  $\Lambda_0$  is chosen such that a maximum allowable false alarm rate (FAR) is achieved (Van Trees, 1968). The likelihood ratio is calculated by:

$$\Lambda(\mathbf{w}) = \frac{p_1(\mathbf{w})}{p_0(\mathbf{w})}. \quad (10.7)$$

To simplify the calculation, we can establish a threshold on the logarithm of the likelihood ratio:

$$\ln \Lambda(\mathbf{w}) = \ln \left[ \frac{p_1(\mathbf{w})}{p_0(\mathbf{w})} \right] = \mathbf{x} \cdot \mathbf{s} - \frac{1}{2}\mathbf{s} \cdot \mathbf{s}. \quad (10.8)$$

A true signal,  $\mathbf{s}$ , has a transit shape that is a function of its host star's effective temperature, mass, radius, and metallicity, as well as the planet's orbital parameters, such as duration and impact parameter. The strength, or depth, of a transit is given by the ratio of the planet area to that of its host star. The orbital parameters of the planet then determine its period,  $T$ , epoch,  $t_0$ , and transit duration,  $d$ . To test for the presence of a transit pulse train, a grid of templates is computed that cover the three-dimensional parameter space of period, epoch, and duration. The transit shape used in TPS is the integral average over the space of all astrophysical transit shapes across all the targets that we observe. All the targets use the same bank of templates for computational efficiency in this first stage of detecting signals. The shape models are built using the Mandel-Agol geometric transit model (Mandel & Agol, 2002) and the limb darkening of Claret & Bloemen (2011). A more complete description of the Monte Carlo used to estimate the integral average is given in Seader et al. (2015). On average, we have the best possible shape match without any *a priori* knowledge of either the target star or any potential planet that may be transiting across it. Choosing a particular point in the  $\{T, t_0\}$  space selects out a set,  $\mathcal{S}$ , of  $P$  samples, marking the center of each transit, starting with the sample corresponding to the epoch  $t_0$  and being spaced  $T$  samples apart. These samples form a subset of  $\{n\}$ ,  $\mathcal{S} = \{t_0, t_0 + T, \dots, t_0 + (P - 1)T\}$ .

The full transit pulse train signal in a template can be written as the sum over pulses (in practice, however, it would be difficult to separate out the inter-pulse correlations that arise from the whitening process discussed above):

$$\mathbf{s} = \sum_{j=1}^P \mathbf{s}_j, \quad (10.9)$$

where  $\mathbf{s}_j$  is a time series vector containing a single transit centered at the  $j$ 'th time in set  $\mathcal{S}$ . The log likelihood ratio can now be rewritten as:

$$\ln \Lambda = \sum_{j=1}^P \left[ \mathbf{x} \cdot \mathbf{s}_j - \frac{1}{2} \mathbf{s}_j \cdot \mathbf{s}_j \right]. \quad (10.10)$$

This can be represented in terms of normalized signal components by letting  $\mathcal{D}^2$  represent the total energy in the signal:

$$\mathcal{D}^2 = \sum_{j=1}^P \mathbf{s}_j \cdot \mathbf{s}_j, \quad (10.11)$$

where the normalized signal components are given by:

$$\hat{\mathbf{s}}_j = \frac{\mathbf{s}_j}{\mathcal{D}} = \frac{\mathbf{s}_j}{\sqrt{\sum_{i=1}^P \mathbf{s}_i \cdot \mathbf{s}_i}}. \quad (10.12)$$

The log likelihood ratio can then be expressed as:

$$\ln \Lambda = \mathcal{D} \sum_{j=1}^P \mathbf{x} \cdot \hat{\mathbf{s}}_j - \frac{1}{2} \mathcal{D}^2. \quad (10.13)$$

To arrive at the final detection statistic, we apply the Maximum Likelihood Method (MLM) and maximize the log likelihood ratio over  $\mathcal{D}$ :

$$\ln \Lambda|_{\hat{\mathcal{D}}} = \frac{1}{2} \left( \sum_{j=1}^P \mathbf{x} \cdot \hat{\mathbf{s}}_j \right)^2, \quad (10.14)$$

where

$$\hat{\mathcal{D}} = \sum_{j=1}^P \mathbf{x} \cdot \hat{\mathbf{s}}_j. \quad (10.15)$$

Here,  $\hat{\mathcal{D}}$  is the value of  $\mathcal{D}$  that maximizes the log likelihood ratio. Rather than thresholding on this, we can equivalently drop the prefactor and the square to arrive at the MES,  $Z$ :

$$Z = \sum_{j=1}^P \hat{\mathbf{s}}_j \cdot \mathbf{x} = \frac{\sum_{j=1}^P \mathbf{s}_j \cdot \mathbf{x}}{\sqrt{\sum_{j=1}^P \mathbf{s}_j \cdot \mathbf{s}_j}}.$$

Note that further analytic maximization of this detection statistic over any of the physical parameters of the true planetary transit signal appears to be intractable due to the functional form of the signal. Therefore, the MES is the Maximum Likelihood detection statistic. If we consider only

a single transit at some time  $n = j$ , the MES reduces to a Single Event Statistic (SES),  $z_j$  given by:

$$z_j = \frac{\mathbf{x} \cdot \mathbf{s}_j}{\sqrt{\mathbf{s}_j \cdot \mathbf{s}_j}}. \quad (10.16)$$

To make explicit the decomposition of the signal into an amplitude, or depth term,  $\mathcal{A}$  and a shape term, we can write (under  $H1$ ):

$$\mathbf{x} = \mathbf{w} + \mathcal{A} \mathbf{s}. \quad (10.17)$$

Under the assumed noise properties (and further assuming no mismatch between the template and the real signal), the relevant statistical properties of the MES are given by:

$$\begin{aligned} \langle Z \rangle &= \mathcal{A} \mathcal{D} \text{ and} \\ \langle Z^2 \rangle &= 1 + \mathcal{A}^2 \mathcal{D}^2, \end{aligned} \quad (10.18)$$

where again,

$$\mathcal{D} = \sqrt{\sum_{j=1}^P \sum_{n=1}^N \tilde{s}_j^2(n)}, \quad (10.19)$$

and the statistical properties under  $H0$  can be obtained by letting  $\mathcal{A} \rightarrow 0$ . Note that the statistical properties of the MES are unchanged under  $H0$  if we allow for mismatch between the signal and template. As we are interested in the distribution of the null statistics, we do not need a requirement that the signal and templates match. The SES and MES are weighted sums of unit-variance Gaussian random variables. Since the sum of squares of the weights add to unity, both the SES and MES are also unit-variance Gaussian random variables. The means of the SES and MES distributions are zero when there is no signal present but are shifted away from zero when a signal is present as can be seen from the above equations. The PDFs of  $Z$  under the two hypotheses are:

$$p_0(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right) \text{ and} \quad (10.20)$$

$$p_1(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(Z - \mathcal{A}\mathcal{D})^2\right). \quad (10.21)$$

Under the Neyman-Pearson criteria, the search threshold,  $\eta$ , is set so that a maximum allowable false alarm rate is achieved. For *Kepler*, the threshold is set to  $7.1\sigma$  to achieve a false alarm rate of  $10^{-12}$  (Jenkins, 2002). This is based on an estimate of the effective number of independent statistical tests in searching the full parameter space for every target that, when multiplied by the maximum allowable false alarm rate, will yield  $\sim 1$  false alarm for the mission (Jenkins, 2002). The false alarm and detection probabilities,  $Q_0$  and  $Q_d$  respectively, are given by:

$$Q_0 = \int_{\eta}^{\infty} p_0(Z) dZ \text{ and} \quad (10.22)$$

$$Q_d = \int_{\eta}^{\infty} p_1(Z) dZ. \quad (10.23)$$

The bootstrap test described in this paper provides a way of estimating the distribution of the null MES, thereby enabling the computation of the false alarm probability for a detection having a given MES and a given number of transits. This is the probability that the underlying noise alone could produce an equivalent detection statistic in the absence of any true signal.

### 10.3 Bootstrap Test

Consider a TCE exhibiting  $p$  transits of a given duration within its light curve. In this analysis, the light curve is viewed as one realization of a stochastic process. Further, consider the collection of all such  $p$ -transit detection statistics that could be generated if we had access to an infinite number of such realizations. To approximate this distribution we formulate the SES time series for the light curve and exclude points that are in-transit (plus some padding) for the given TCE. Bootstrap statistics are then generated by drawing at random from the SES  $p$  times, with replacement, to formulate the  $p$ -transit statistic (i.e., the MES). Since the SES encapsulate the effects of local correlations in the background noise process on the detectability of transits, so does each individual bootstrap statistic. So long as the orbital period is sufficiently long (generally longer than several hours), the SES are uncorrelated and the MES can be considered to be formed by  $p$  independent random deviates from the distribution of null SES.

The MES distribution is a complicated function of the noise as well as the epoch, period, and transit duration. Due to deviations from the assumed noise behavior, the MES distribution can be significantly different across epochs, periods, or even transit durations, so it is generally not appropriate to simply sample over the full parameter space to estimate the MES distribution. The bootstrap method presented in Jenkins (2002) sorts the data in such a way as to minimize the computation time in estimating the tail end of the distribution needed for computing the false alarm rate. This method proved intractable from a computational standpoint when  $p \geq 15$ . The method presented here attempts to estimate the distribution for a given period and transit duration. In this way any noise behavior that deviates from expectation on the same time scale of the detection will be encoded in the distribution to give a more reliable estimate of the false alarm probability localized in parameter space.

The MES,  $Z$ , can be rewritten as:

$$Z = \frac{\sum_{j=1}^P \mathbf{s}_j \cdot \mathbf{x}}{\sqrt{\sum_{j=1}^P \mathbf{s}_j \cdot \mathbf{s}_j}} = \sum_{i \in \mathcal{S}} \mathbb{C}(i) \Bigg/ \sqrt{\sum_{i \in \mathcal{S}} \mathbb{N}(i)}, \quad (10.24)$$

where  $\mathcal{S}$  is the set of transit times that a single period and epoch pair select out,  $\mathbb{C}(i)$  is the correlation time series formed by correlating the whitened data to a whitened transit signal template with a transit centered at the  $i^{th}$  timestep in the set  $\mathcal{S}$ , and  $\mathbb{N}(i)$  is the template normalization time series. The square root of the normalization time series,  $\sqrt{\mathbb{N}(i)}$ , is the expected value of the MES or SNR for the reference transit pulse.<sup>4</sup>

If the observation noise process underlying the light curve is well modeled as a possibly non-white, possibly non-stationary Gaussian noise process, then the SES will be zero-mean, unit-variance (ZMUV) Gaussian random deviates. The false alarm rate of the transit detector would then be described by the complementary distribution for a zero-mean, unit-variance (ZMUV) Gaussian distribution:

$$\bar{F}_Z(Z) = \frac{1}{2} \operatorname{erfc}\left(Z/\sqrt{2}\right), \quad (10.25)$$

where  $\operatorname{erfc}(\cdot)$  is the standard complementary error function. If the power spectral density of the noise process is not perfectly captured by the whitener in TPS, then the null statistics will not be zero-mean, unit-variance Gaussian deviates. A bootstrap analysis allows us to obtain a data-driven approximation of the actual distribution of the null statistics, rather than relying on the assumption that the pre-whitener is perfect.

<sup>4</sup>The inverse of  $\sqrt{\mathbb{N}(i)}$  can be interpreted as the effective white Gaussian noise “seen” by the reference transit and is the definition for the combined differential photometric precision (CDPP) reported for the *Kepler* light curves at 3, 6, and 12 hours’ duration.

The random variable  $Z$  is a function of the random variables corresponding to the correlation and normalization terms in the SES time series  $\mathbb{C}_p = \sum_{i \in S} \mathbb{C}(i)$  and  $\mathbb{N}_p = \sum_{i \in S} \mathbb{N}(i)$ . The joint density of  $\mathbb{C}_p$  and  $\mathbb{N}_p$  can be determined from the joint density of the SES components  $\mathbb{C}$  and  $\mathbb{N}$  as:

$$f_{\mathbb{C}_p, \mathbb{N}_p}(\mathbb{C}_p, \mathbb{N}_p) = f_{\mathbb{C}, \mathbb{N}}(\mathbb{C}, \mathbb{N}) * f_{\mathbb{C}, \mathbb{N}}(\mathbb{C}, \mathbb{N}) * \dots * f_{\mathbb{C}, \mathbb{N}}(\mathbb{C}, \mathbb{N}), \quad (10.26)$$

where ‘\*’ is the convolution operator and the convolution is performed  $p$  times. This follows from the fact that the bootstrap samples are constructed from *independent* draws from the set of null SES with replacement.<sup>5</sup> Given that convolution in the time/spatial domain corresponds to multiplication in the Fourier domain, Equation 10.26 can be represented in the Fourier domain as:

$$\Phi_{\mathbb{C}_p, \mathbb{N}_p} = \Phi_{\mathbb{C}, \mathbb{N}} \cdot \Phi_{\mathbb{C}, \mathbb{N}} \cdot \dots \cdot \Phi_{\mathbb{C}, \mathbb{N}} = \Phi_{\mathbb{C}, \mathbb{N}}^p, \quad (10.27)$$

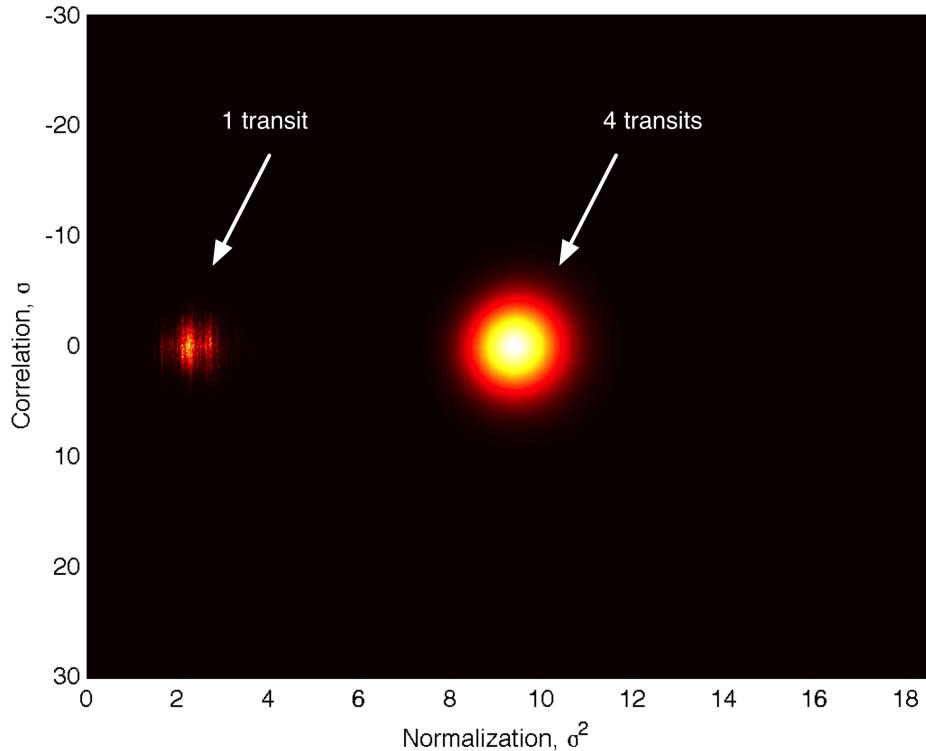
where  $\Phi_{\mathbb{C}, \mathbb{N}} = \mathcal{F}\{f_{\mathbb{C}, \mathbb{N}}\}$  is the Fourier transform of the joint density function  $f_{\mathbb{C}, \mathbb{N}}$ . Here, the arguments of the Fourier transforms of the density functions have been suppressed for clarity. The use of 2-D fast Fourier transforms results in a highly tractable algorithm from a computational point of view. Figure 10.2 illustrates the construction of the MES distribution for the case of a 4-transit TCE.

The implementation of Equation 10.27 requires that a 2-D histogram be constructed for the  $\{\mathbb{C}, \mathbb{N}\}$  pairs over the set of null statistics. Care must be taken to manage the size of the histogram to avoid spatial aliasing as the use of the FFT corresponds to circular convolution. We chose to formulate the 2-D grid to allow for as high as  $p = 8$  transits. The intervals covered by the realizations (i.e., the support) for each of  $\mathbb{C}$  and  $\mathbb{N}$  were sampled with 256 bins and centered in a 4096 by 4096 array. When  $p > 8$ , it is necessary to implement Equation 10.27 iteratively in stages; after each stage the characteristic function is transformed back into the spatial domain, bin-averaged by a factor of 2, and then padded back out to the original array size. Care must also be taken to manage knowledge of the zero-point of the histogram in light of the circular convolution, as it shifts by a 1/2 sample with each convolution operation in each dimension.

Once the  $p$ -transit 2-D density function  $f_{\mathbb{C}_p, \mathbb{N}_p}(\mathbb{C}_p, \mathbb{N}_p)$  is obtained, it can be “collapsed” into the sought-after 1-D density  $f_Z(Z)$  by mapping the sample density for each cell with center coordinates  $\{\mathbb{C}_i, \mathbb{N}_j\}$  to the corresponding coordinate  $Z_{i,j} = \mathbb{C}_i / \sqrt{\mathbb{N}_j}$  and formulating a histogram with a resolution of, say,  $0.1 \sigma$  in  $Z$  by summing the resulting densities that map into the same bins in  $Z$ . Due to the use of the FFT, the precision of the resulting density function is limited to the floating point precision of the variables and computations, which is  $\sim 2.2 \times 10^{-16}$ . For small  $p$ , the density may not reach to the limiting numerical precision because of small number statistics, and for large  $p$ , round-off errors can accumulate below about  $10^{-14}$ . The results can be extrapolated to high MES values by fitting the mean,  $\mu$ , and standard deviation,  $\sigma$ , of a Gaussian distribution to the empirical distribution in the region  $10^{-4} \leq \bar{F}_Z \leq 10^{-13}$  using the standard complementary error function:  $\bar{F}_Z(Z) = 0.5 \operatorname{erfc}((Z - \mu) / \sqrt{2} \sigma)$ . Note that the fitted distribution can only be used to extrapolate the upper tail of the bootstrap distribution and is not valid for describing the core of the empirical distribution, as it is not constrained to fit the latter below  $10^{-4}$ .

In order to simulate the use of  $\chi^2$ -square vetoes (Seader et al., 2013) that effectively remove strong transient and impulsive features that trigger TPS but that are inconsistent with physical transit signatures, we pre-filtered the SES time series to remove the three most positive peaks and their “shoulders” down to  $2 \sigma$ . The three most negative peaks were also handled in a similar

<sup>5</sup>In this implementation, we choose to assume that the null statistics are governed by a single distribution. If this is not the case (for example, if the null statistic densities vary from quarter to quarter) then Equation 10.26 could be modified to account for the disparity in the relevant SES in a straightforward manner.



**Figure 10.2** Illustration of the bootstrap algorithm and Equation 10.27. False color image of the correlation component  $\mathbb{C}$  versus the normalization component  $\mathbb{N}$  for the bivariate distributions for the single event null statistics and for the 4-transit multiple event statistic distribution for one TCE. Although the 1-transit distribution is highly irregular, that of the 4-transit distribution is much more symmetric and Gaussian, as follows from the central limit theorem.

fashion to avoid biasing the mean of the null statistics in a negative direction. We also identified and removed points with a density of zero-crossings that fell below 1/4 that of the median zero-crossing density. This step removed SES in regions where the correlation term experienced strong excursions from zero due to unmitigated SPSDs and thermal transients near monthly and quarterly boundaries. Typically, more than 99% of the original out-of-transit SES were retained by these pre-filters.

#### 10.4 Archive Column Definitions

There are four quantities derived from the bootstrap test that are present in the TCE table. These four quantities are defined as follows:

`boot_fap`: The false alarm probability (FAP), which is the integral of the distribution of the null MES above the MES of the detection. The distribution of the null MES is constructed by the bootstrap test. Nominally, the null MES is Gaussian distributed with zero mean and unit variance. In reality however, due to imperfections in the whitening process, uncorrected systematics, etc., the distribution of the null MES deviates from this nominal distribution form.

`boot_mesthresh`: The search threshold required, given the distribution of the null MES estimated from the bootstrap algorithm, to achieve the same false alarm probability as that of a  $7.1\sigma$  threshold on a Gaussian distribution with zero mean and unit variance ( $\bar{F}_Z \sim 6.24 \times 10^{-13}$ ).

`boot_mesmean`: The mean of the best-fit Gaussian distribution to the upper tail ( $10^{-13} \leq \bar{F}_Z \leq 10^{-4}$ ) of the null MES distribution estimated by the bootstrap. This quantity, together with the quantity `boot_messtd`, is useful for extrapolating the false alarm probability to values less than  $10^{-13}$  and for high MES values.

`boot_messtd`: The standard deviation of the best-fit Gaussian distribution to the upper tail ( $10^{-13} \leq \bar{F}_Z \leq 10^{-4}$ ) of the null MES distribution estimated by the bootstrap.

In cases where there is not enough data to run the bootstrap, the false alarm probability is set to -1. In some cases, the bootstrap test cannot use the data from the distribution it has constructed to interpolate for the false alarm probability; rather, it must extrapolate because the MES is outside the regime that the distribution covers. To do the extrapolation, a robust fit of an error function is done in log space to the Cumulative Distribution Function (CDF) of the MES with  $10^{-13} \leq \bar{F}_Z \leq 10^{-4}$ . The parameters of the fit are used to calculate the false alarm probability for the MES of the detection. Features in the CDF can sometimes cause the fit to be poor, which in turn causes the fit parameters and resulting false alarm probabilities to also be poor.

## 10.5 Results

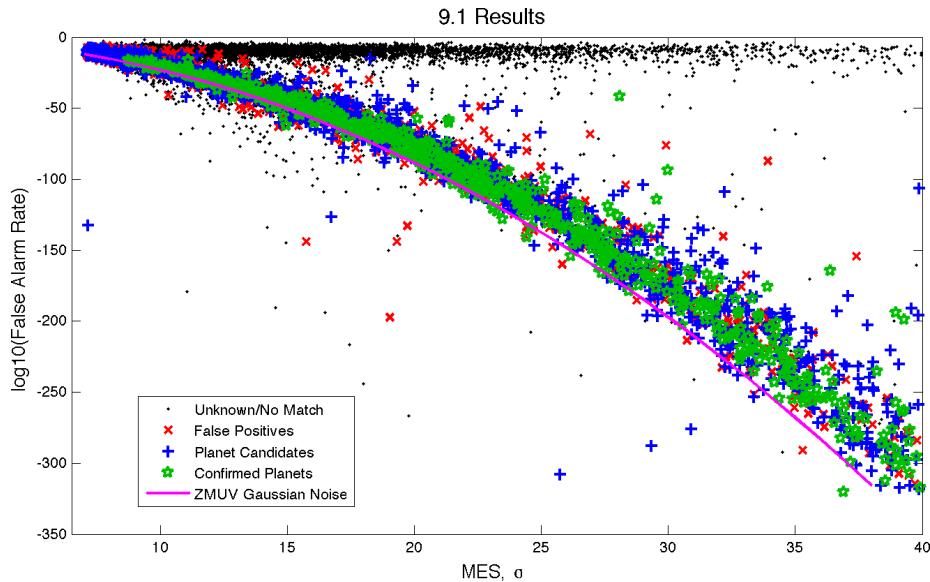
The last few years have seen significant improvements in the SOC science data processing pipeline, leading to higher quality light curves and more sensitive transit searches. The statistical bootstrap analysis results presented here and the numerical results archived at NASA's Exoplanet Science Institute (NExSci) bear witness to these software improvements. This section attempts to introduce and describe the main features and differences between these three datasets as a consequence of the software changes.

### 10.5.1 Q1–Q16

Figure 10.3 and Figure 10.4 show the false alarm probability for 16,014 TCEs as a function of the MES value of each TCE. A small fraction of the TCEs' light curves had too few points remaining after removing in-transit and neighboring points to conduct a bootstrap analysis. We matched the Q1–Q16 TCEs against the cumulative KOI table through Data Release 24 (Coughlin et al., 2016) on NExSci's exoplanet archive. The KOIs (consisting of a mixture of planet candidates, confirmed/validated planets and astrophysical false positives) tend to lie in a band whose left edge is approximately enveloped by the curve expected for ZMUV Gaussian noise. At low SNR, there is a small population of planet candidates and astrophysical false positives with false alarm rates above  $10^{-10}$  that are embedded in a much larger population of TCEs that form a horizontal “cloud” where the false alarm rate is nearly independent of the MES. Visual inspection of the light curves underlying points in this “cloud” indicate that most of these light curves are polluted with residual stellar variability and flares. Many appear consistent with being red giants with power spectral densities richly populated by pressure mode oscillations. The whitener in TPS is not designed to handle such noise, resulting in spurious detections.

Most of the points falling below the ZMUV curve are very short-period TCEs with 500 or more transits. These targets have little data left after the removal of in-transit samples, and the remaining single event null statistics are slightly biased with a mean below zero. For periods sufficiently short such that there are  $\sim$ 500 or more transits, computing the bootstrap distribution for the MES involves raising the 2-D characteristic function for the single event null statistics to the number of observed transits (see Equation 10.27). This process yields a distribution with a mean that is significantly negative. This results in false alarm rates well below those expected for ZMUV noise for such cases.

Figure 10.5 shows a plot of the false alarm rate as a function of the bootstrap threshold (`boot_mesthresh`) for the SOC 9.1 Q1–Q16 TCEs, colored by disposition for the KOIs



**Figure 10.3** False alarm rate as a function of the multiple event statistic (MES) for each of the 16,014 TCEs returning bootstrap results in the Q1–Q16 transiting planet search. The points are colored by the dispositions of the TCEs in NExSci Exoplanet Archive’s cumulative KOI table through DR24. The magenta line indicates the expected value for a zero-mean, unit-variance (ZMUV) Gaussian process.

matched against the TCE ephemerides.<sup>6</sup> Note that some confirmed/validated planets and planet candidates have thresholds above  $\sim 10 \sigma$ . The confirmed planets include Kepler-90d and h, Kepler-30c and d, and Kepler-444e and f. All of these are systems exhibit strong transit timing variations and therefore the use of a linear ephemeris to mask out transits leaves residual transits in the null statistic time series, inflating the bootstrap thresholds and false alarm rate estimates for these cases.

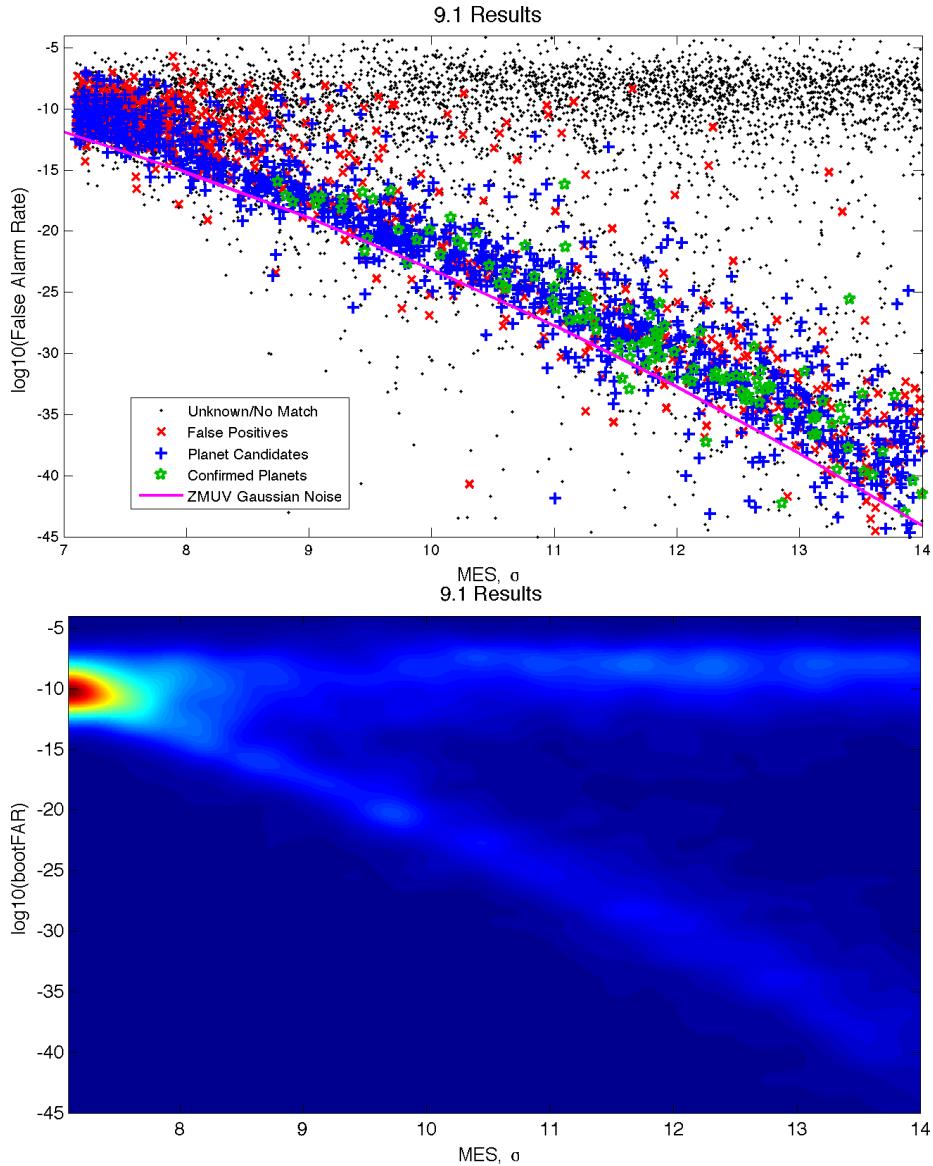
We note that the bootstrap threshold (`boot_mesthresh`) is negative for a few stars. This is most often the case for very short orbital periods ( $\leq 5$  days) where the residual SES are so cut up by the removal of transit signatures that they are biased negative relative to the true mean and the resulting N-transit statistics are significantly skewed. This is a limitation of the methodology, but is expected behavior. The `boot_mesmean` can be negative as well because the upper tail of the empirical bootstrap distribution is being modeled and the fit is unconstrained. The model is only useful for extrapolating to MES values above that observed in the empirical distribution where the false alarm rate is  $\ll 10^{-4}$ . If a valid fit cannot be obtained for some reason, then the Gaussian fit parameters are gapped (i.e., marked as unpopulated).

While the bootstrap results for the well-behaved transit signatures and the spurious detections in the horizontal cloud are not well separated at low SNR ( $< 9\sigma$ ), the bootstrap false alarm probabilities are strong indicators for high SNR TCEs. Improvements in the quality of the light curves and the TPS codebase dramatically improve the situation, as will be seen in Subsection 10.5.3.

### 10.5.2 SOC 9.2 Q1–Q17 DR24

The SOC 9.3 bootstrap algorithm was run on the SOC 9.2 Q1–Q17 DR24 TCEs previously and archived at NExSci (see KSCI-19086-003). That set of bootstrap results only included quarters

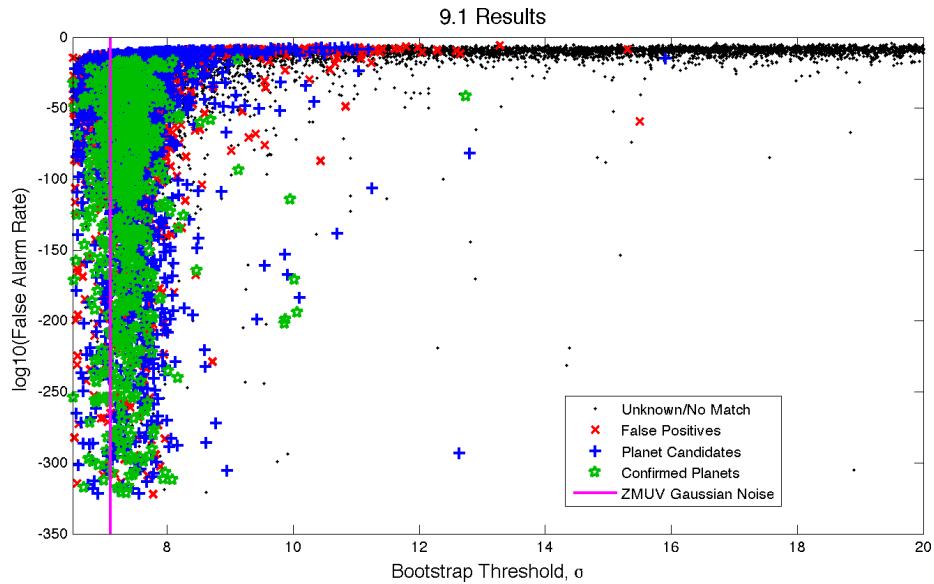
<sup>6</sup>For this document, a match was considered valid if both the epoch and the period of the TCE were within 0.1 days of the given KOI.



**Figure 10.4** Zoomed view of Figure 10.3. Top panel: False alarm rate as a function of the MES for each of the Q1–Q16 TCEs. Bottom panel: Density plot of the false alarm rate as a function of the MES. Note that the two principal populations, the horizontal branch with little dependency on MES and the one that is approximately enveloped by the expected curve for ZMUV Gaussian noise, both merge at low SNR ( $< 9\sigma$ ) near  $\log(\text{FAR}) = 10^{-10}$ .

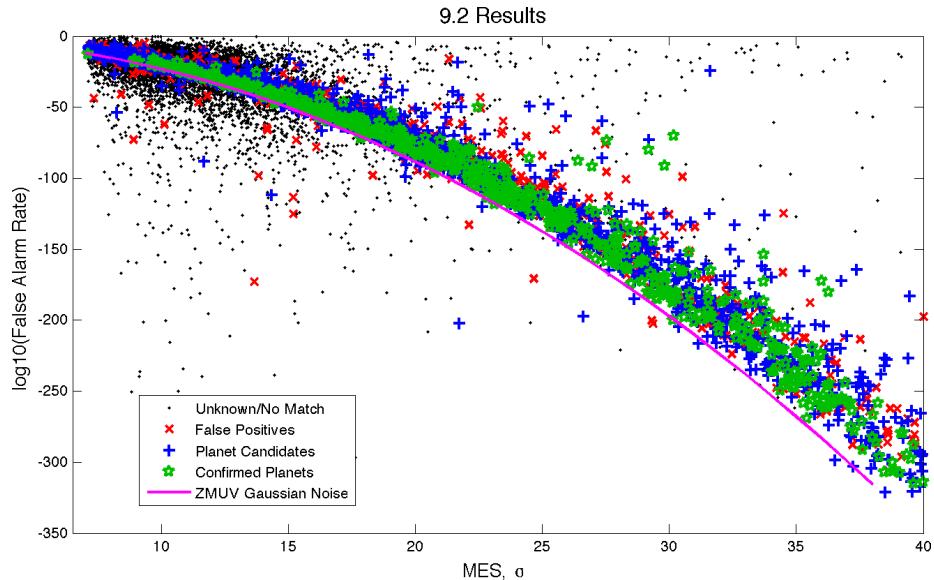
with transits in the calculation. This note documents redelivery of the bootstrap results for this dataset using all available data from Q1–Q17 in the bootstrap analysis, for consistency with the other two datasets delivered at this time.

The results are similar to those for the Q1–Q16 dataset, although since the SOC 9.2 bootstrap was used as a veto in TPS for this run, the horizontal “cloud” of high MES/high false alarm rate objects is missing, as almost all of these objects were rejected in TPS and not presented to DV for further analysis and characterization (see Christiansen et al. (2016) for a detailed discussion of the impact of the bootstrap veto on completeness of the transit search). Figure 10.6 and



**Figure 10.5** False alarm rate as a function of the bootstrap threshold for the Q1–Q16 TCEs. For the KOIs recovered in this dataset, the points are colored by the disposition in the cumulative DR24 KOI table. The vertical line at  $7.1 \sigma$  represents the threshold expected for ZMUV Gaussian noise.

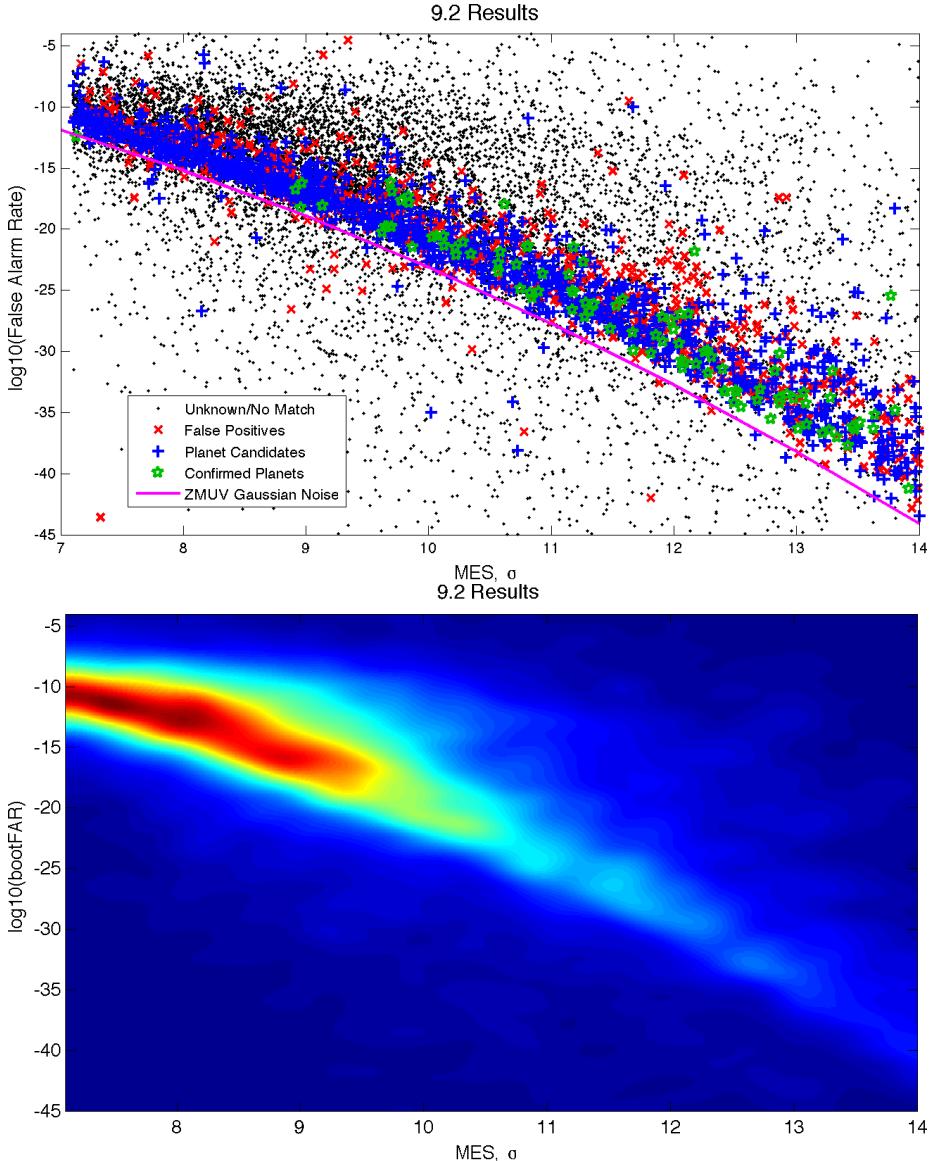
Figure 10.7 present the bootstrap false alarm rate as a function of the MES, while Figure 10.8 presents the bootstrap false alarm rate as a function of the bootstrap threshold.



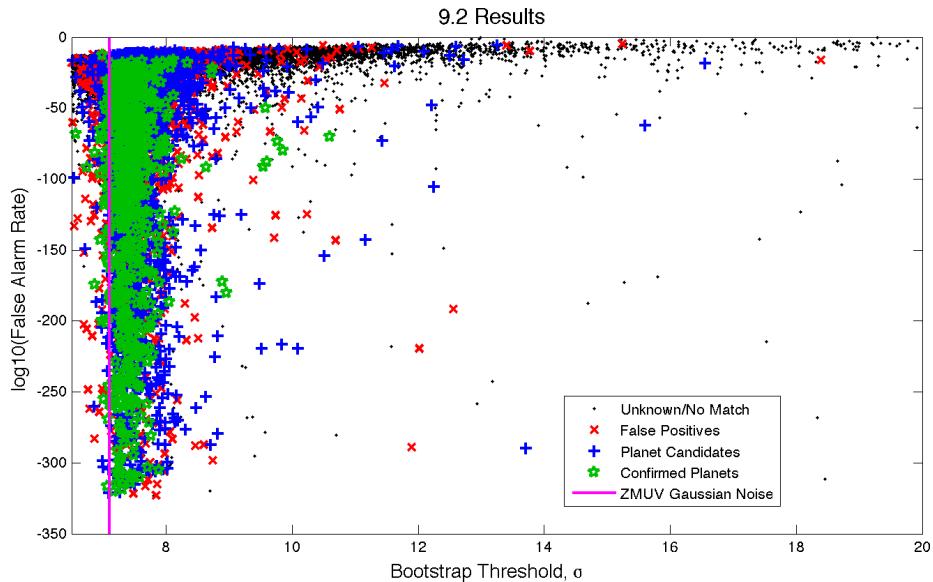
**Figure 10.6** False alarm rate as a function of the MES for each of the 19,856 TCEs returning bootstrap results in the Q1–Q17 DR24 transiting planet search. The points are colored by KOI disposition. The magenta line indicates the expected value for a ZMUV Gaussian process.

Because the horizontal “cloud” feature is completely absent from Figure 10.6, we can conclude that the bootstrap is quite effective at filtering non-transit-like features associated with this

population from the TPS results. The KOIs identified in the SOC 9.2 DR24 TCEs fall along a band that is enveloped on the left by that expected for ZMUV Gaussian noise. As with the SOC 9.1 results, none of the confirmed or validated planets have bootstrap false alarm rates above  $10^{-12}$ , suggesting that the bootstrap can be used to screen against spurious TCEs, although it is unclear whether this can be done without rejecting true transiting planets at low SNR ( $<9\sigma$ ) given that the large population of spurious TCEs evident for SOC 9.1 are not present in this SOC 9.2 dataset.



**Figure 10.7** Zoom of Figure 10.6. Top panel: False alarm rate vs. MES for the 19,856 SOC 9.2 Q1–Q17 DR24 TCEs. Bottom panel: Density plot of the false alarm rate as a function of the multiple event statistic. Note that the horizontal branch with little dependency on MES is missing for SOC 9.2 as the bootstrap was used as a veto in TPS.



**Figure 10.8** False alarm rate as a function of the bootstrap threshold for the 19,856 Q1–Q17 DR24 TCEs, colored by KOI disposition. The vertical line at  $7.1\sigma$  represents the threshold expected for ZMUV Gaussian noise.

### 10.5.3 SOC 9.3 Q1–Q17 DR25

Figure 10.9 and Figure 10.10 present the false alarm rate as a function of MES for the 24,179 TCEs returning bootstrap results in the SOC 9.3 Q1–Q17 DR25 search. It is interesting to note that the ZMUV curve is a much better envelope for the band of TCEs that are also KOIs. Note also that this band of KOIs is tighter than those for either of the two previous datasets. This is largely due to changes made in the SOC 9.3 TPS codebase to improve the performance of the whitening filter as well as changes in the assignment of photometric apertures (Smith et al., 2016) and improvements in systematic error correction. These changes and their effect on the DR25 TCEs are documented in Twicken et al. (2016).

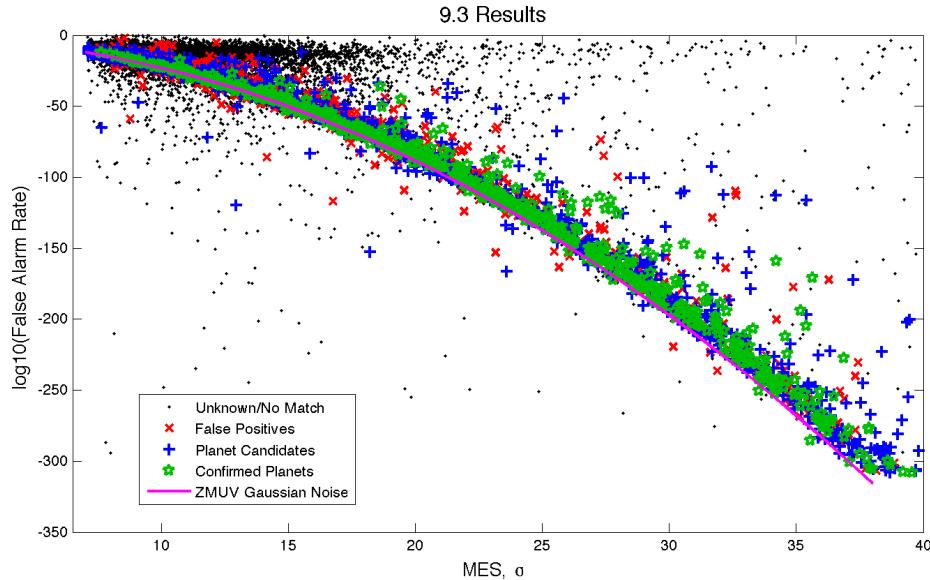
One change in the SOC 9.3 TPS codebase was to incorporate a non-decimated moving median absolute deviation (MAD) filter for estimating the RMS noise power as a function of time in each of the wavelet filter bank’s bandpasses. (These noise power estimates are used to implement the adaptive whitener in the matched filter search for transiting planets.)

This change was motivated by the observation that there was a measurable and significant duration-dependent bias in the noise power estimates. The noise power for short-duration transits was underestimated relative to that for long-duration transits. Prior to SOC 9.3, TPS made use of a decimated moving MAD filter due to computational throughput constraints. We were able to recode the moving MAD filter algorithm in C++ and eliminate the bias in the noise power estimates while maintaining adequate computational throughput. This specific change significantly reduced the vertical spread in the bootstrap false alarm rates for the KOIs.

Another change was to use a sigmoid taper for filling long gaps in the flux time series, which decreased edge effects in the SES and greatly reduced the “droop” in the noise power estimates in the neighborhood of these long gaps (see Chapter 9 for more information).

For the SOC 9.3 DR25 results, there is much better separation of the low-reliability TCEs in the horizontal “cloud” population from those in the band following the ZMUV curve compared to the SOC 9.1 TCEs. The unreliable TCEs have false alarm probabilities greater than approx-

imately  $10^{-11}$  while all confirmed and/or validated planets, almost all planet candidates, and most astrophysical false positives have  $\log_{10}(\text{FAR}) < 10^{-11}$ .



**Figure 10.9** False alarm rate as a function of the multiple event statistic for each of the 24,179 TCEs returning bootstrap results in the SOC 9.3 Q1–Q17 DR25 transiting planet search. The points are colored by the dispositions of the TCEs in NExScI Exoplanet Archive’s cumulative DR24 KOI table. The magenta line indicates the expected value for a ZMUV Gaussian process.

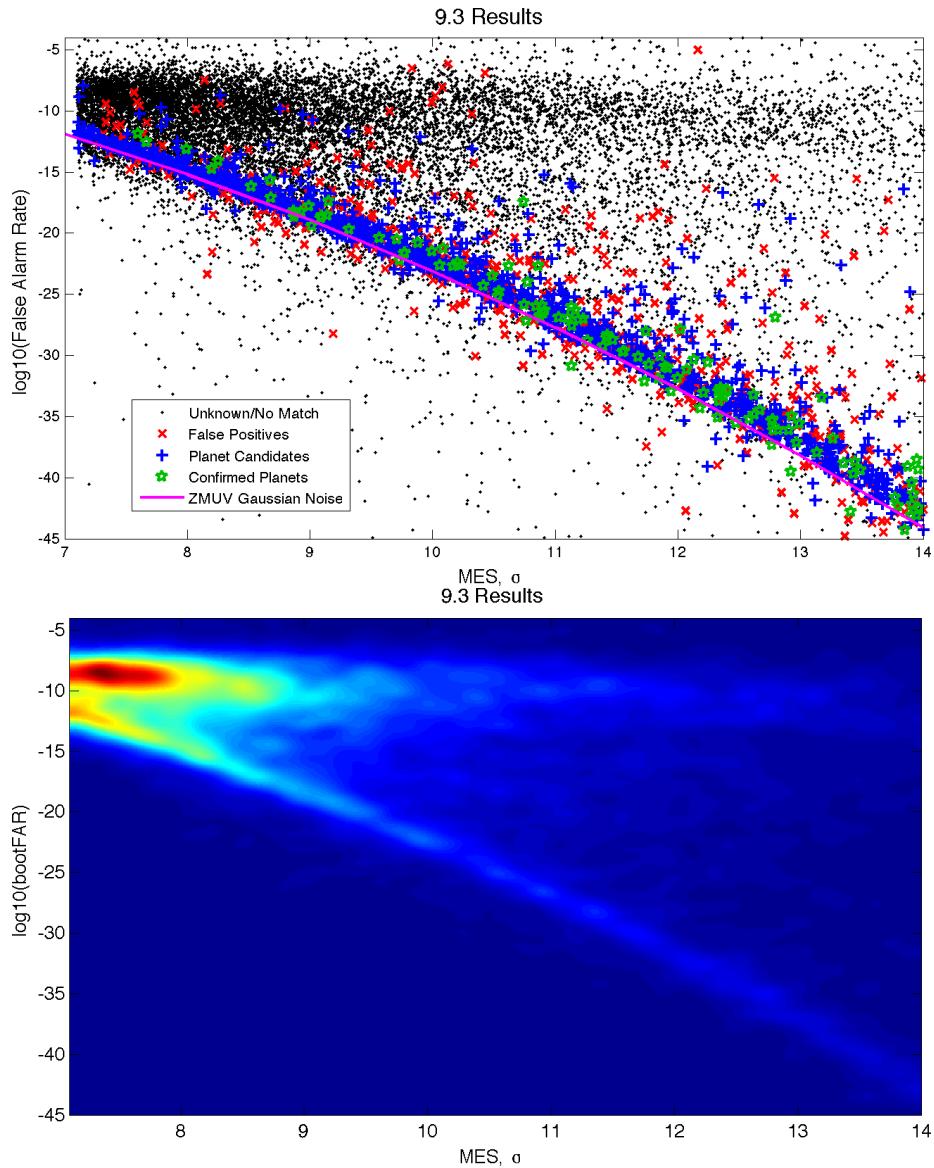
Figure 10.11 shows the bootstrap false alarm rate as function of bootstrap threshold. The SOC 9.3 changes also reduced the number of planet candidates with bootstrap thresholds  $>10\sigma$  from 46 in SOC 9.1 (see Figure 10.5) and 56 in SOC 9.2 (see Figure 10.8) to only 13 in SOC 9.3.

#### 10.5.4 Comparison Across Datasets

While the results of the bootstrap analysis are similar across the three datasets, there are differences due in small part to the amount of data considered in each one (16 quarters of data vs. 17 quarters of data), and in large part to code base changes for the production of the light curves. In this section we discuss some of the differences between the datasets.

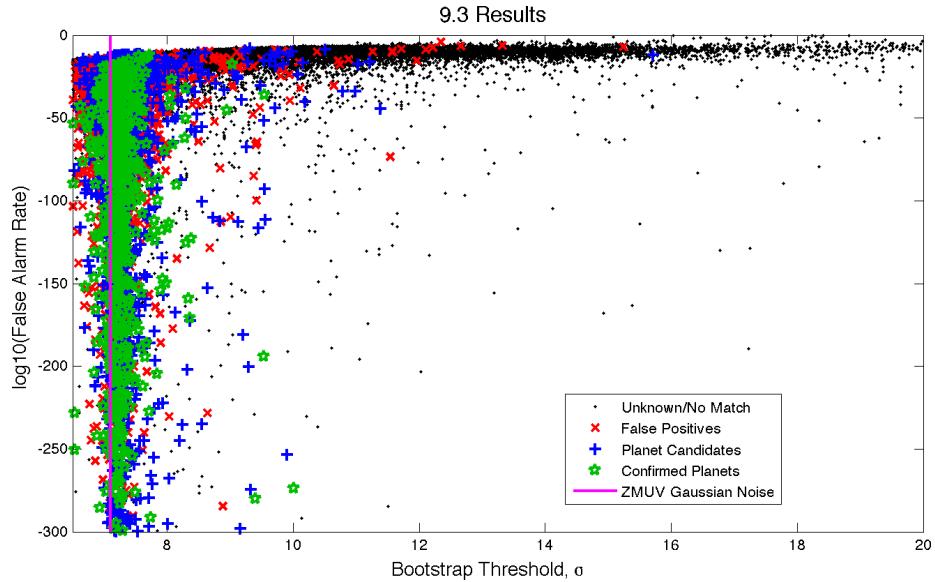
Figure 10.12 shows the difference in the log of the bootstrap false alarm rates between SOC 9.2 and SOC 9.1, colored by KOI disposition, where the 9.2 false alarm rates are calculated using the SOC 9.1 MES values to ensure that the comparisons are valid. For false alarm rates exceeding  $\sim 10^{-40}$  there is a relatively tight core of points between  $\sim 10^{1.6}$  and  $\sim 10^{-0.6}$  for the KOIs. The points are well correlated between the two datasets considering the differences in the codebases and the fact that results smaller than  $\sim 10^{-16}$  are all extrapolations from fits to the upper tails of the empirical bootstrap distributions.

Figure 10.13 shows the difference in the log of the bootstrap false alarm rates between SOC 9.3 and SOC 9.1, colored by KOI disposition, where the 9.3 false alarm rates are calculated using the SOC 9.1 MES values to ensure that the comparisons are valid. For false alarm rates greater than  $\sim 10^{-40}$  there is a relatively tight core of points between  $\sim 10^{0.5}$  and  $\sim 10^{-2}$  for the confirmed/validated planets, planetary candidates, and astrophysical false positives. The SOC 9.3 results are somewhat lower than those for the same objects for SOC 9.1, reflecting the improvements in the flux time series and in the TPS algorithm due to the SOC 9.3 codebase changes.

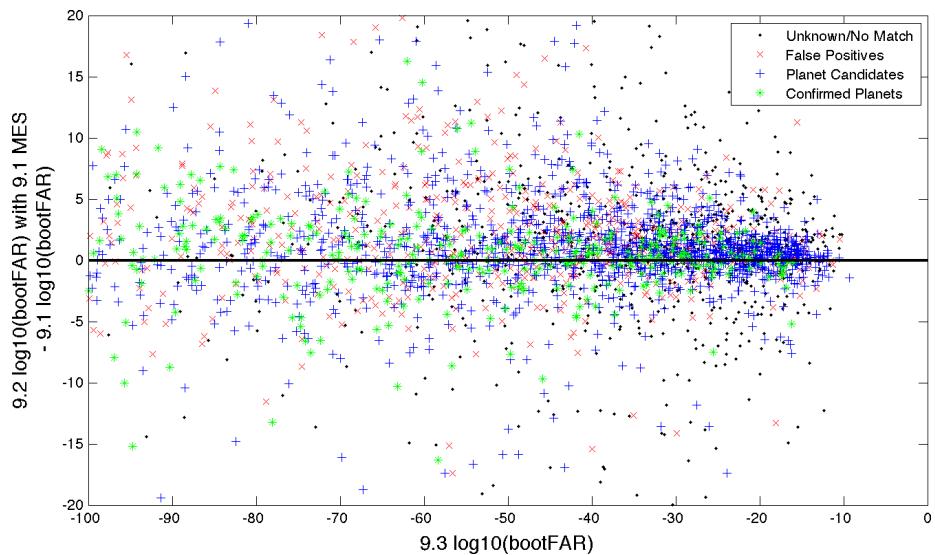


**Figure 10.10** Zoom of Figure 10.9. Top panel: False alarm rate as a function of the MES for each of the SOC 9.3 Q1–Q17 DR25 TCEs, colored by KOI disposition. Note that the “cloud” of high MES/high FAR reappears here as the bootstrap was *not* used as a veto in TPS for this run. Bottom panel: Density plot of the false alarm rate as a function of the MES. Note that the two principal populations, the horizontal branch with little dependency on MES and the one that is approximately enveloped by the expected curve for ZMUV Gaussian noise are well separated in the SOC 9.3 results with a valley between them at  $\log(\text{FAR}) \approx 10^{-11}$ .

The codebase changes from SOC 9.1 to SOC 9.3 also tightened the distribution of bootstrap thresholds produced by the statistical bootstrap analysis. In addition, the median threshold dropped from  $7.44\sigma$  in SOC 9.1 to  $7.24\sigma$  in SOC 9.3, demonstrating the increase in sensitivity as the SOC codebase evolved. The 10th, 50th, and 90th percentiles for the bootstrap thresholds for TCEs that were matched against confirmed/validated planets and astrophysical false positives are given in Table I.

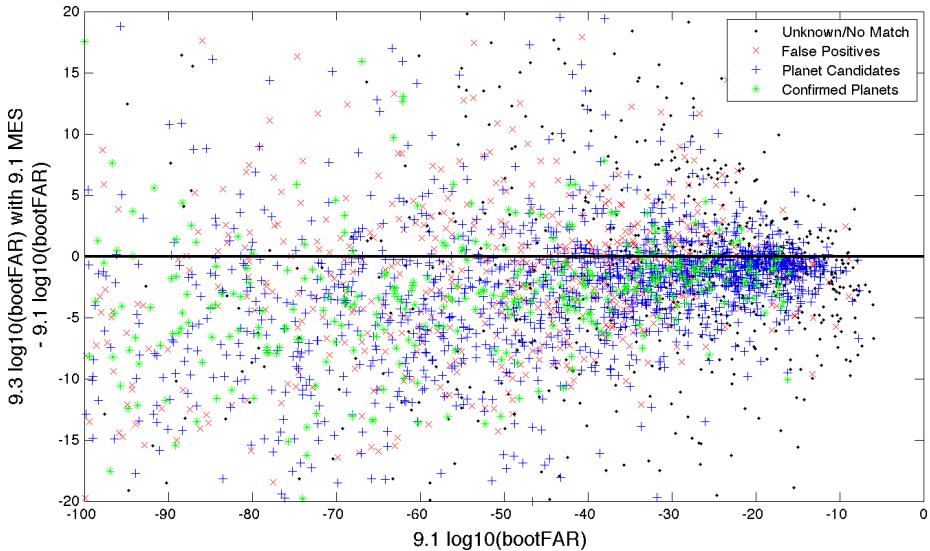


**Figure 10.11** False alarm rate as a function of the multiple event statistic for each of the 24,179 TCEs returning bootstrap results in the SOC 9.3 Q1–Q17 DR25 transiting planet search, colored by the KOI disposition. The magenta line indicates the expected value for a ZMUV Gaussian process.



**Figure 10.12** Difference in false alarm rate for the SOC 9.2 Q1–Q17 TCEs vs. that for the SOC 9.1 Q1–Q16 TCEs as a function of the SOC 9.1 false alarm rates for those objects that have matching ephemerides in the two datasets. The SOC 9.2 false alarm rates are calculated for the SOC 9.1 MES values to ensure a valid comparison. The SOC 9.2 false alarm rates are slightly higher than those for SOC 9.1, perhaps due to differences in the codebases.

Figure 10.14 shows histograms of the bootstrap thresholds for KOIs for each of the datasets. The SOC 9.3 results show a broad upper tail, thanks to improved sensitivity in the search and relaxation of the vetoes in TPS for the final run, which doubled the number of TCEs from 16,285 in SOC 9.1 to 34,032 in SOC 9.3.



**Figure 10.13** Difference in false alarm rate for the SOC 9.3 Q1–Q17 TCEs vs. that for the SOC 9.1 Q1–Q16 TCEs as a function of the SOC 9.1 false alarm rates for those objects that have matching ephemerides in the two datasets. The SOC 9.3 false alarm rates are calculated for the SOC 9.1 MES values to ensure a valid comparison. The SOC 9.3 false alarm rates are somewhat lower than those for SOC 9.1 due to the improvements in the photometric pipeline as well as in TPS.

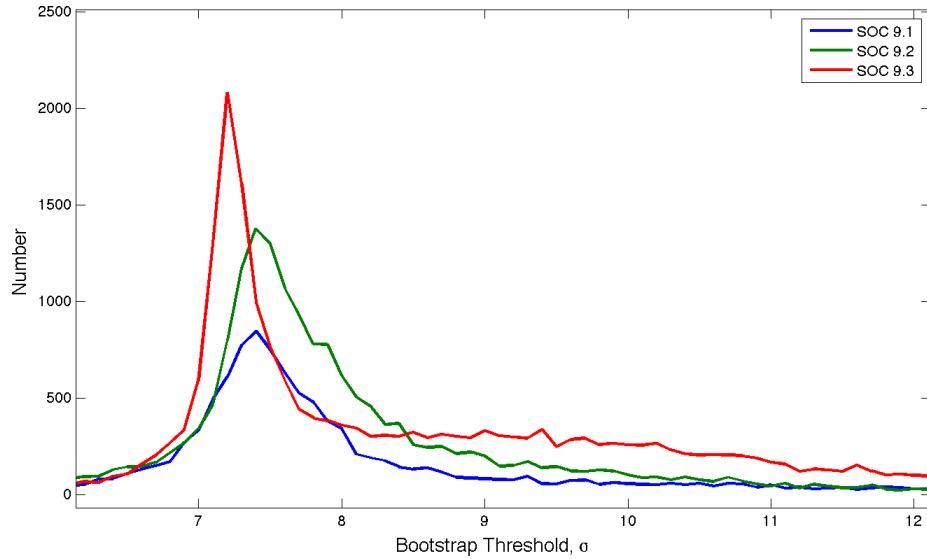
**Table 10.1** Summary Statistics of the Bootstrap Thresholds for KOIs.

Release	10th Percentile	50th Percentile	90th Percentile
SOC 9.1	$6.84 \sigma$	$7.44 \sigma$	$8.36 \sigma$
SOC 9.2	$7.17 \sigma$	$7.51 \sigma$	$8.36 \sigma$
SOC 9.3	$7.02 \sigma$	$7.24 \sigma$	$7.75 \sigma$

## 10.6 Precision of the Statistical Bootstrap Results

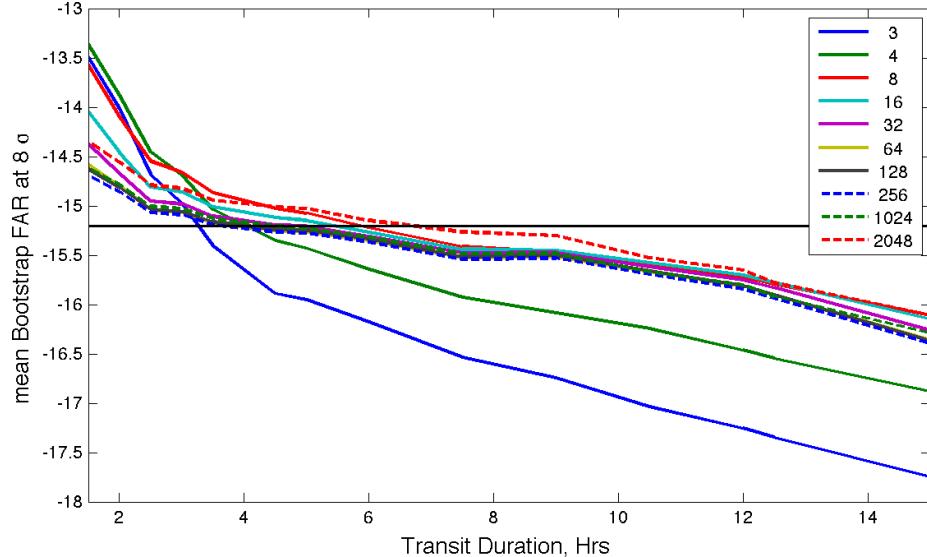
In this section we investigate the precision of the statistical bootstrap test by reviewing the results of a Monte Carlo experiment. The light curve for a typical Kepler target observed for all 17 quarters was replaced with a ZMUV WGN process and run through TPS to generate the SES time series. This process incorporated all gaps in the original time series in order to present as realistic a result as possible. The SES time series for each of the 14 different pulse durations searched in TPS (between 1.5 and 15 hours) were then subjected to the SOC 9.3 bootstrap analysis. The number of transits,  $n_{\text{transits}}$ , for each pulse duration was varied between 3 and 2048. A total of 100 random flux time series were generated and subjected to the bootstrap analysis in this manner.

Figure 10.15 shows the bootstrap false alarm rate at  $8 \sigma$  as a function of the transit duration and the number of transits. For transit pulse durations less than  $\sim 3$  hours, the false alarm rate is slightly higher than that expected for a fully sampled ZMUV Gaussian process, which would provide a  $\log_{10}(\text{FAR})$  of  $-15.2$ . The FAR then drops gradually to the longest duration, perhaps reflecting the fact that there are more independent statistical tests conducted for shorter transits than for longer-duration transits (at a given trial orbital period). The two smallest number of transit cases also exhibit depressed false alarm rates as a function of duration. For eight or more



**Figure 10.14** Bootstrap thresholds for KOIs for SOC 9.1 (blue curve), SOC 9.2 (green curve) and SOC 9.3 (red curve).

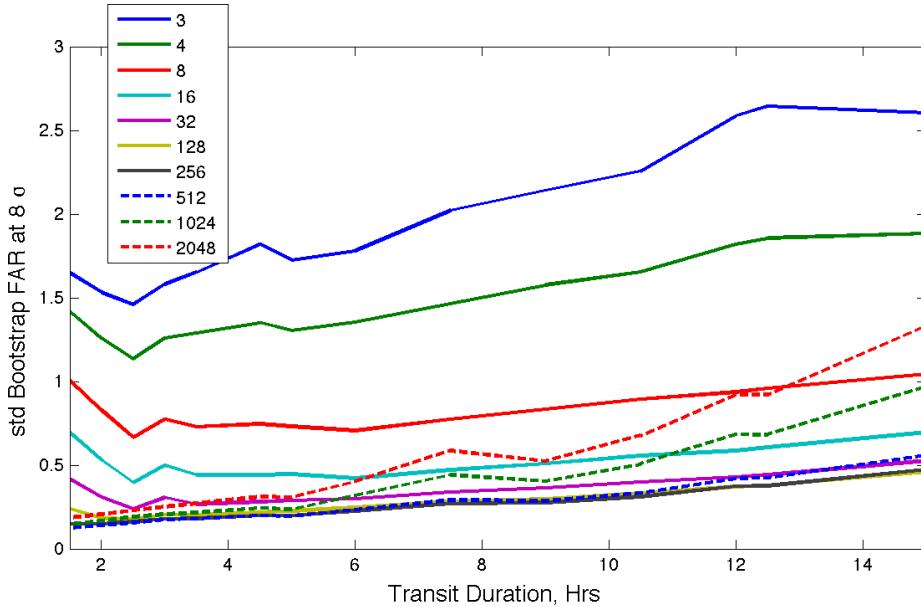
transits, the FAR at  $8\sigma$  converges to a rather narrow curve with a width of  $\sim 0.25$  in log space. For transit durations exceeding 3 hours the dispersion is 2 dex.



**Figure 10.15** Mean of the log of the false alarm rate at  $8\sigma$  of 100 different Monte Carlo tests as a function of transit duration and number of transits. The solid black line indicates the idealized false alarm rate at  $8\sigma$ , namely  $\log_{10}(6.2 \times 10^{-16}) = -15.206$ .

Figure 10.16 shows the standard deviation in the  $\log_{10}(\text{FAR})$  at  $8\sigma$  across all 100 Monte Carlo runs as a function of transit duration and number of transits. The curve for  $n_{\text{transits}}=3$  is between 1.5 and 2.6 over the full range of durations. The curves drop rapidly as a function of  $n_{\text{transits}}$ . For  $n_{\text{transits}} \geq 8$  the standard deviation of the bootstrap FAR at  $8\sigma$  is less than 1 dex. For  $n_{\text{transits}} \geq 512$ , the scatter in the results begins increasing, reflecting accumulation of round

off errors due to the spatial resampling that must be used in the bootstrap analysis to prevent spatial aliasing (see Section 10.3).



**Figure 10.16** Standard deviation of the log of the false alarm rate at  $8\sigma$  of 100 different Monte Carlo tests as a function of transit duration and number of transits. The solid line indicates the ideal threshold at  $7.1\sigma$ .

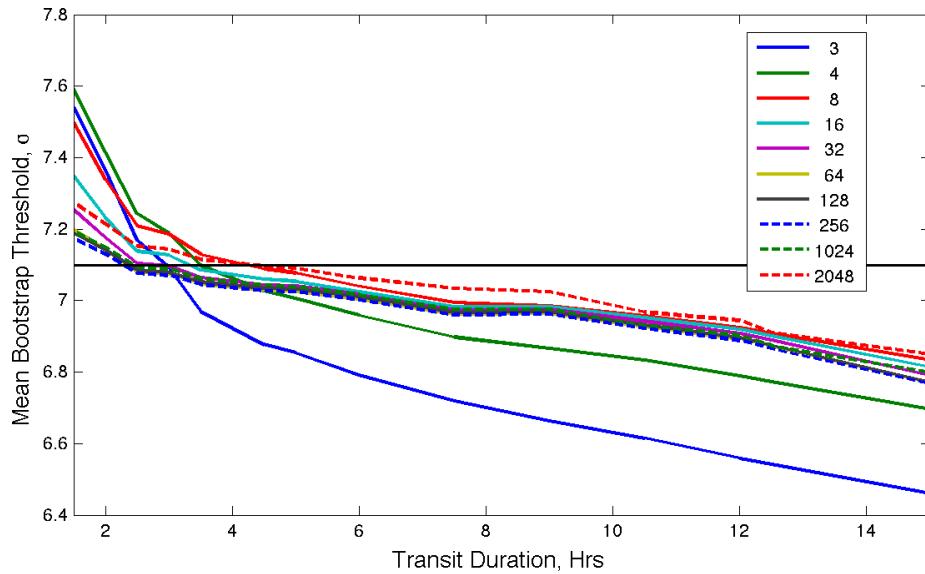
Note that the dispersion in the mean bootstrap FAR is comparable to the scatter in Monte Carlo results, indicating that while the mean bootstrap FAR does indeed vary significantly with respect to transit duration and orbital period (number of transits), the bias in an individual bootstrap FAR estimate is approximately the same as the uncertainty for long orbital periods ( $n_{\text{transits}} < 8$ ) and long durations ( $> 6$  hr). For these cases, the bootstrap FAR is likely to be biased by 1–2 dex towards smaller false alarm rates.

Figure 10.17 shows the mean of the bootstrap threshold as a function of transit duration and  $n_{\text{transits}}$ . The shape of the curves are rather similar to those for the mean  $\log_{10}(\text{FAR})$  in Figure 10.15. For durations longer than 3 hours and  $n_{\text{transits}} \geq 8$  the bootstrap threshold drops linearly from  $7.1\sigma$  to  $6.6\sigma$ . The curves for  $n_{\text{transits}} \geq 8$  are relatively tightly confined to within a range of  $\pm 0.03\sigma$ . The variation across the curves is  $\sim 0.4\sigma$  at any given transit pulse duration, comparable to the scatter of the results across the Monte Carlo trials.

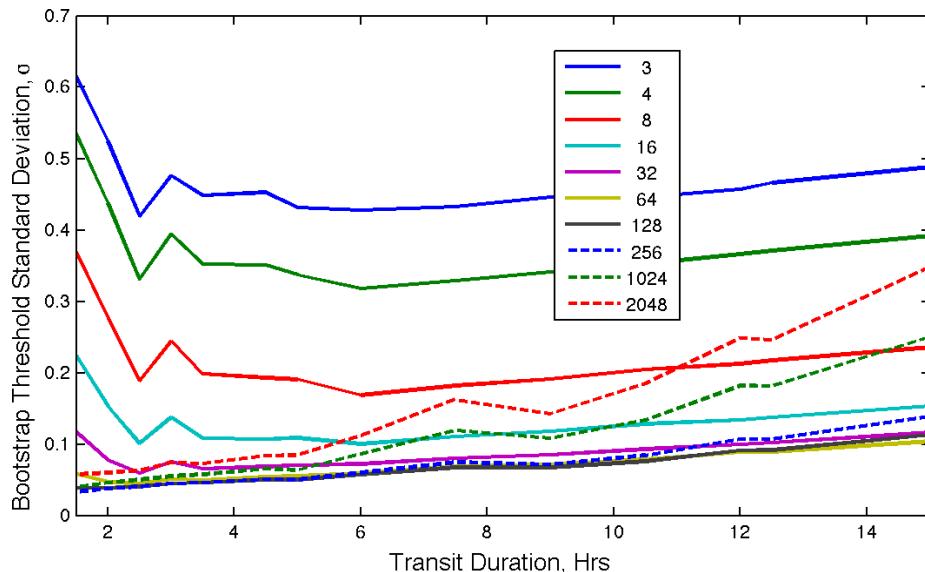
Figure 10.18 shows the standard deviation of the bootstrap threshold as a function of the transit duration and  $n_{\text{transits}}$ . The shape of the curves is reminiscent of those for the standard deviation of the  $\log_{10}(\text{FAR})$  in Figure 10.15. The standard deviation of the threshold is below  $0.5\sigma$  for all cases, dropping rapidly from  $\sim 0.5\sigma$  for  $n_{\text{transits}} = 3$  to  $\sim 0.2\sigma$  for  $n_{\text{transits}} = 8$ .

The results are well behaved for transit pulse durations  $\geq 3$  hours, and for  $\geq 8$  transits. However, the enhanced scatter in the results for small numbers of transits should be kept in mind when interpreting the bootstrap results for TCEs with  $n_{\text{transits}} < 8$ .

How much of the bias structure evident in Figure 10.15 and Figure 10.17 are due to the bootstrap, and how much is due to the conditioning, filtering, and processing within TPS? We conducted a separate Monte Carlo experiment by running bivariate Gaussian MES through the bootstrap algorithm directly, thereby bypassing TPS. Figure 10.19 shows the behavior of the mean bootstrap FAR and threshold at  $8\sigma$  as a function of the number of transits for this experiment along with those for the original Monte Carlo experiments. The  $\log_{10}(\text{FAR})$  of the bivariate

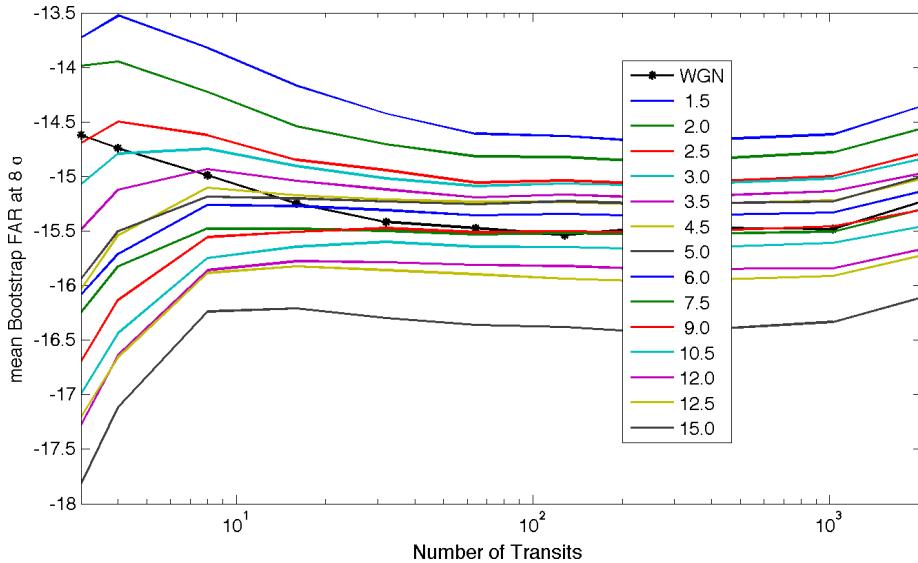


**Figure 10.17** Mean of the bootstrap threshold of 100 different Monte Carlo tests as a function of transit duration and number of transits.



**Figure 10.18** Standard deviation of the bootstrap threshold of 100 different Monte Carlo tests as a function of transit duration and number of transits.

WGN process varies between -14.5 and -15.4, or  $\pm 1$  dex, indicating that the majority of the bias structure dependent on transit duration is due to the filtering and conditioning occurring inside of TPS. We interpret this as being due to the fact that the shorter-duration transits have less data “averaged” into each SES, and hence, are noisier than those for longer duration transits, and to the fact that for a finite flux time series, there are more effective independent statistics for shorter duration transits relative to longer duration transits.



**Figure 10.19** Mean of the bootstrap FAR of 100 different Monte Carlo tests as a function of transit duration and number of transits, along with that for a bivariate WGN process passed directly to the bootstrap algorithm (black curve with stars).

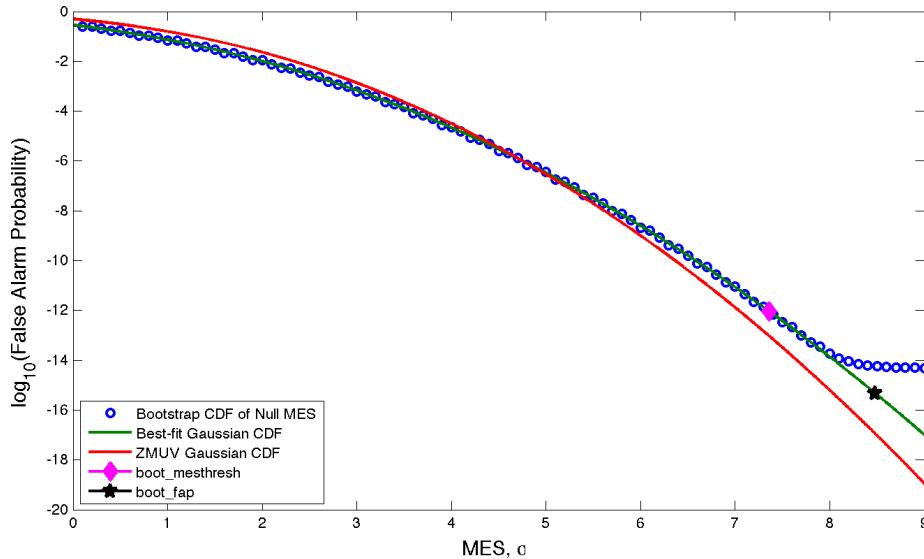
Note that while these Monte Carlo experiments give some idea of the native scatter in the bootstrap analysis results, they do not include all the known instrumental artifacts (e.g. sudden pixel dropouts or rolling band image artifacts) and/or astrophysical red noise.

## 10.7 Bootstrap Analysis of a Single TCE

As an illustration of how the statistical bootstrap analysis operates, Figure 10.20 shows a typical bootstrap result. The TCE used for this plot is on KIC 12158032, which has a transit duration of 2 hours, an orbital period of 0.578 days, and a MES of  $8.48\sigma$ . If the MES of the detection falls above the MES corresponding to a  $\log_{10}(\text{FAP})$  of -13.5, where FAP is the false alarm probability, then the `boot_fap` is interpolated from the CDF of the null MES constructed by the bootstrap, otherwise the best-fit Gaussian is used to calculate the `boot_fap`. In Figure 10.20 it is marked by the black star and was calculated to be  $4.7 \times 10^{-16}$ . This is the FAP on the solid green curve corresponding to the MES of the TCE. The `boot_mesthresh` for this TCE is  $\sim 7.35\sigma$  as can be seen by finding the MES corresponding to a FAP of  $\sim 6.24 \times 10^{-13}$  on the best-fit Gaussian (indicated by the magenta diamond). The `boot_mesmean` is the mean of the best-fit Gaussian and is -0.64 for this TCE. The `boot_messtd` is the standard deviation of the best-fit Gaussian and is 1.13 for this TCE. Note that the solid red curve shows the CDF for a ZMUV Gaussian. The Gaussian is fitted robustly in log space using the data from  $1 \times 10^{-4}$  to  $1 \times 10^{-13}$  to avoid the roll-off toward  $\text{MES} > 8\sigma$  due to round-off errors.

## 10.8 Conclusions

This chapter provided the motivation and mathematical development for the statistical bootstrap analysis conducted in DV for each TCE identified by TPS. We introduced a new algorithm that is computationally efficient and able to provide bootstrap results for all TCEs regardless of how many transits participate in the transit signature. The bootstrap results for each of the past three



**Figure 10.20** The CDF of the null MES constructed by the bootstrap. This TCE is on KIC 12158032 and has a MES of  $8.48\sigma$ , a duration of 2 hours, and a period of 0.578 days and yielded 2,278 transits in 4 years of data. The false alarm probability for this TCE is  $\sim 4.7 \times 10^{-16}$ , marked by the black star. The best-fit Gaussian has a mean of -0.64 and a standard deviation of 1.13. The magenta diamond marks the threshold needed to achieve the same false alarm rate of a ZMUV Gaussian with a  $7.1\sigma$  threshold given the distribution of null MES constructed by the bootstrap.

transit searches, including the final one over all 17 quarters of data with the final codebase, SOC 9.3, known as Data Release 25, were described and compared. A set of Monte Carlo experiments were performed to assess the statistical precision of the bootstrap products available to the scientific community on the NExSci exoplanet archive. Finally, a numerical example was provided for a single TCE.

## Bibliography

- Borucki, W. J., Koch, D., Basri, G., et al., 2010. “Kepler Planet-Detection Mission: Introduction and First Results,” *Science*, 327, 977
- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al., 2016. “Measuring Transit Signal Recovery in the Kepler Pipeline. III. Completeness of the Q1-Q17 DR24 Planet Candidate Catalogue with Important Caveats for Occurrence Rate Calculations,” *ApJ*, 828, 99
- Claret, A., & Bloemen, S., 2011. “Gravity and Limb-Darkening Coefficients for the *Kepler*, CoRoT, Spitzer, uvby, UBVRJHK, and Sloan photometric systems,” *Astronomy & Astrophysics*, 529, A75, available online at <http://dx.doi.org/10.1051/0004-6361/201116451>
- Coughlin, J. L., Mullally, F., Thompson, S. E., et al., 2016. “Planetary Candidates Observed by Kepler. VII. The First Fully Uniform Catalog Based on the Entire 48-month Data Set (Q1-Q17 DR24),” *ApJS*, 224, 12
- Jenkins, J. M., 2002. “The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets,” *ApJ*, 575, 493

- Jenkins, J. M., Caldwell, D. A., & Borucki, W. J., 2002. “Some Tests to Establish Confidence in Planets Discovered by Transit Photometry,” *ApJ*, 564, 495
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010. “Transiting Planet Search in the Kepler Pipeline,” in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400D
- Koch, D. G., Borucki, W. J., Basri, G., et al., 2010. “Kepler Mission Design, Realized Photometric Performance, and Early Science,” *ApJL*, 713, L79
- Mandel, K., & Agol, E., 2002. “Analytic Light Curves for Planetary Transit Searches,” *ApJL*, 580, L171
- Seader, S., Tenenbaum, P., Jenkins, J. M., & Burke, C. J., 2013. “ $\chi^2$  Discriminators for Transiting Planet Detection in Kepler Data,” *ApJS*, 206, 25
- Seader, S., Jenkins, J. M., Tenenbaum, P., et al., 2015. “Detection of Potential Transit Signals in 17 Quarters of Kepler Mission Data,” *ApJS*, 217, 18
- Smith, J. C., Morris, R. L., Jenkins, J. M., et al., 2016. “Finding Optimal Apertures in Kepler Data,” *PASP*, 128, 124501
- Tenenbaum, P., Christiansen, J. L., Jenkins, J. M., et al., 2012. “Detection of Potential Transit Signals in the First Three Quarters of Kepler Mission Data,” *ApJS*, 199, 24
- Tenenbaum, P., Jenkins, J. M., Seader, S., et al., 2014. “Detection of Potential Transit Signals in 16 Quarters of Kepler Mission Data,” *ApJS*, 211, 6
- Twicken, J. D., Jenkins, J. M., Seader, S. E., et al., 2016. “Detection of Potential Transit Signals in 17 Quarters of Kepler Data: Results of the Final Kepler Mission Transiting Planet Search (DR25),” *AJ*, 152, 158
- Van Trees, H. L. 1968, *Detection, Estimation, and Modulation Theory, Part I* (Wiley), 19–155, 239–442



## CHAPTER 11

---

# DATA VALIDATION I – ARCHITECTURE, DIAGNOSTIC TESTS, AND DATA PRODUCTS

**Abstract.** A Threshold Crossing Event is generated in the Transiting Planet Search (TPS) component of the Pipeline for each target with a transit-like signature which exceeds the transit detection threshold. These targets are subjected to an automated validation process in the Data Validation (DV) component of the Pipeline. Transiting planet candidates are characterized in DV, and light curves are searched for additional transiting planets after transit signatures are modeled and removed until further candidates are no longer identified (or an iteration limit is reached). A suite of diagnostic tests is subsequently performed on all candidates to aid in discrimination between true transiting planets and instrumental or astrophysical false positives. Data products are generated per target and planet candidate to document and display the planet model fit and diagnostic test results. These products are exported to the Exoplanet Archive at the NASA Exoplanet Science Institute and are available to the community at large for vetting transiting planet candidates identified in the *Kepler* Science Data Processing Pipeline. We describe the DV architecture and diagnostic tests in this chapter, and provide a brief overview of the DV data products. Transiting planet modeling and the search for multiple planets on individual targets are described subsequently in Chapter 12.

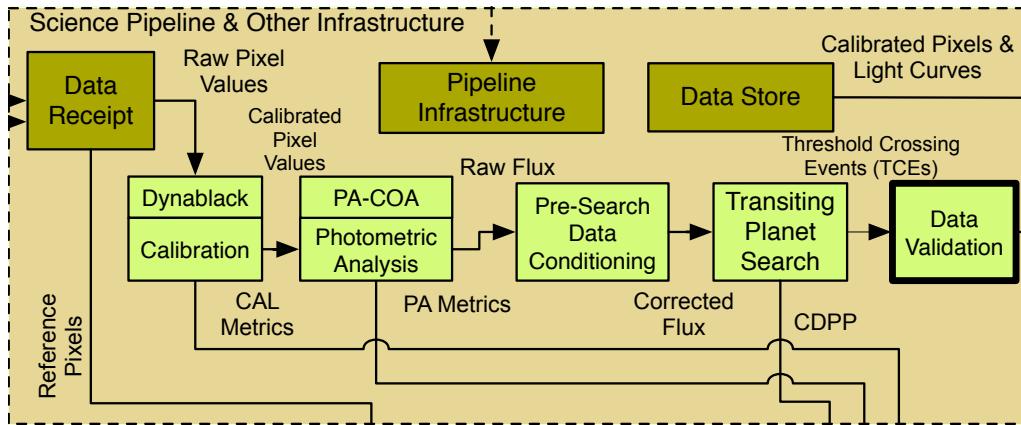
**Keywords:** Stars; Extrasolar Planets; Characterization; Validation; Data Analysis and Techniques

### 11.1 Introduction

Please note that this chapter is in a preliminary state and will be updated in a revised version of the *Kepler* Data Processing Handbook as soon as it is available.

This chapter discusses the architecture of the Data Validation (DV) module of the *Kepler* Science Processing Pipeline and details the diagnostic tests applied to each transit-like feature, or threshold crossing event (TCE), identified by the Transiting Planet Search (TPS) module. Chapter 12 documents the fitting of a limb-darkened transit model to each TCE before constructing the diagnostic tests documented in this chapter. Figure 11.1 shows the DV module in the context of the *Kepler* Science Data Processing Pipeline.

The transit model fitting is designed for the following two main tasks: 1) The fitted parameters of the transit model and the corresponding light curve generated from the model are used in the diagnostic tests in DV to disposition planetary candidates; 2) When the TPS module is called,



**Figure 11.1** Data Validation (DV) in the context of the architecture of the *Kepler* Data Processing Pipeline. As a part of the DV module, the transit model fitting fits TCEs generated by TPS to derive parameters that are used in various diagnostic tests of DV.

only one TCE with the maximum event detection statistic is generated. To search for multiple-planetary candidates, an iterative process of transit model fitting and multiple-planet search is implemented in DV. For each target star, the transit model parameters are fitted to each TCE generated by TPS, the signature of known TCEs is removed from the light curve, the flux time series, and then the residual goes through TPS again to search for additional TCEs. This iteration will only terminate once no TCEs are identified or a preset upper limit is reached.

The results of the transit model fitting of the TCEs generated by the *Kepler* Data Processing Pipeline, such as the fitted parameters and uncertainties, derived parameters and uncertainties, fit goodness metrics and the diagnostic plots, are included in the DV report or DV one-page summary report, which are both accessible by the science community at NASA Exoplanet Archive (Akeson et al., 2013).<sup>1</sup>

This chapter is organized as follows. Subsection 11.1.1 describes the TCE vetting process and how DV products contribute to that process. Section 11.2 presents a high level view of the DV pipeline module. Section 11.3 introduces and describes the principal diagnostic tests conducted on each TCE, including the weak secondary test (Subsection 11.3.1), the rolling band contamination flags (Subsection 11.3.2), the eclipsing binary discrimination tests (Subsection 11.3.3), the centroid motion test (Subsection 11.3.6), difference imaging (Subsection 11.3.4) and centroid offset analysis (Subsubsection 11.3.4.2), the statistical bootstrap test (Subsection 11.3.5, the centroid motion test (Subsection 11.3.6 ), and the ghost diagnostic test (Subsection 11.3.7). Section 11.5 describes the archive products, including the DV Report (Subsection 11.5.1) and the DV Summary Report (Subsubsection 11.5.1.1). A summary and conclusions are presented in Section 11.6.

### 11.1.1 Vetting Transiting Planet Candidates

Planet candidates are characterized in DV by transiting planet model fitting. Light curves are searched for additional transiting planets after transit signatures are modeled and removed until further candidates can no longer be identified (or an iteration limit is reached). A suite of diagnostic tests is performed on each candidate to aid in discrimination between true transiting planets and instrumental or astrophysical false positives. Data products are generated per target

<sup>1</sup>(<http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=tce>

and planet candidate to document and display the planet model fit and diagnostic test results. These products are exported to the NASA Exoplanet Archive (Akeson et al., 2013) and are available to the community at large for vetting transiting planet candidates identified in the *Kepler* Pipeline. The initial revision of DV was documented by Wu et al. (2010), but the DV component evolved significantly since the time of that publication. The state of the *Kepler* Pipeline Data Validation art is described both in this chapter and its companion, Chapter 12.

The design goals of DV were to (1) characterize planet candidates identified in the Pipeline, and (2) perform powerful diagnostic tests uniformly on all TCEs to aid in assessment of the planet candidates. DV was specifically not tasked with rating, ranking or otherwise classifying Pipeline planet candidates as to the likelihood that they represent true transiting planets. Nor was DV tasked with scoring the value of candidates under the assumption that they represent real planet detections, i.e., an Earth-size planet in the HZ of a Sun-like star is worth far more than a hot Jupiter detectable from the ground from the standpoint of the *Kepler Mission*. Decisions concerning the veracity of the Pipeline candidates and their relative priority were to be left to human experts.

And so it is that the data products generated by DV are employed by TCERT in a multiple stage vetting process. The initial step involves a TCE triage whereby all Pipeline candidates that cannot realistically represent bona fide transit signatures are rejected; such candidates involve instrument and spacecraft anomaly driven TCEs, stellar variability on time scales of transits, and low SNR signatures that would not be credible statistically even if attributable to real transiting planets. Once implausible TCEs are rejected, the remaining TCEs are federated with the list of known KOIs. New candidates are promoted to KOI status. The TCEs associated with new and existing KOIs are then further assessed and classified as PC or FP based on diagnostics derived from light curves, pixel time series, and centroids.

The TCERT vetting process was largely manual well into the primary *Kepler* mission [Refs]. The high cost of the vetting process (in both time and resources) and the reliance on human decision makers subject to individual biases and prone to inconsistency and error led to the development of a rules-based system (“robo-vetter”) for assessing TCEs, promoting worthy Pipeline detections to KOI status, and classifying TCEs associated with KOIs in a given TPS/DV run [Refs]. The Q1–Q17 DR24 and DR25 planet catalogs will both be based on the TCERT robo-vetter.

At the same time, a machine learning system (“auto-vetter”) was developed to employ attributes generated in TPS/DV to classify TCEs generated in a Pipeline run as Planet Candidate, Astrophysical False Positive, or Junk. Classifications are determined by a random forest of decision trees [Refs]. Decision trees are trained with labeled TCEs and then applied to classify unknown (i.e., unlabeled) TCEs based on their respective attributes. Training labels for the auto-vetter are determined in part from prior TCERT vetting activities. The random forest methodology is robust against errors in labeling training data [Ref] and as a byproduct permits the computation of *a posteriori* probabilities for the TCE classifications [Ref].

## 11.2 Pipeline Data Validation

All targets for which a TCE is generated in TPS are processed independently in DV. The architecture of the DV Pipeline software component is shown in Fig. ???. The first step in DV is to characterize transiting planets identified in the Pipeline. Transiting planet modeling is described in detail in Chapter 12 and will only be summarized here. Following preliminary preprocessing steps, a transiting planet model is robustly fitted to the systematic error corrected light curve of the given target. The fitting is performed in a whitened domain where transformed data samples are independent and identically distributed after removal of the transit signature. The whitening is implemented in an adaptive time varying fashion with wavelet-based machinery (Jenkins,

2002, and see Chapter 9). Stitching of the quarterly data into a single contiguous time series and filling of gaps in the available data has been documented with regard to the transiting planet search in Chapter 9; the code for quarter stitching and gap filling is shared between TPS and DV.

The transiting planet model fit is seeded with the TCE period, epoch, trial transit pulse duration, and associated detection statistic. Five parameters are fitted in the model: period, epoch, impact parameter, reduced planet radius ( $R_p/R_*$ ), and reduced semimajor axis ( $a/R_*$ ). After the fit has converged, the fitted transits are removed and the residual light curve is searched for the presence of another transiting planet signature. The search is performed by calling TPS directly from DV. The transiting planet model is fitted to each of the candidates identified in the so-called “multiple planet search.” The process is terminated when no additional TCE is produced in the multiple planet search or a configurable iteration limit is reached. Historically, the iteration limit for the multiple planet search has been set to ten TCEs for any given target. Some targets have produced ten TCEs, but no target has yielded ten credible transit signatures, so the limit has not led to the loss of true planet candidates.

The transiting planet model is also fitted separately to the sequences of odd and even transits for each planet candidate identified in the Pipeline in support of DV diagnostic tests that will be discussed later in this chapter. DV may also be configured to optionally perform a series of “reduced-parameter” fits in which the impact parameter is fixed at specified values while the remaining four model parameters are fitted.

The Mandel-Agol (Mandel & Agol, 2002) transiting planet model is employed to render light curves at the barycentric corrected cadence timestamps (Thompson et al., 2016) in and near transit. A small body approximation is employed to reduce the model run-time when the reduced planet radius is less a specified threshold (typically 0.01). Nonlinear limb darkening coefficients are interpolated from tables produced by Claret & Bloemen (2011) based on stellar parameters for each given target. Stellar parameters provided to DV may be obtained from the *Kepler* Input Catalog (KIC) (Brown et al., 2011) or they may represent overrides to KIC parameters. The KIC overrides for the Q1–Q17 DR25 run were produced by the *Kepler* Stellar Properties Working Group (Huber et al., 2014). Stellar parameters employed in DV are stellar radius, effective temperature, surface gravity ( $\log g$ ) and metallicity (Fe/H). DV assumes Solar values for stellar parameters in cases where LC targets are unclassified in the KIC (i.e., stellar parameters are unspecified) and target-specific overrides are not provided to DV. Provenance is tracked so that the source of stellar parameter values may be documented in the DV archive data products on a parameter by parameter basis.

The following orbital and planet characteristics are derived from the fit parameters after the transiting planet model fits converge: orbital semimajor axis, planet radius, equilibrium temperature, effective stellar flux (i.e., insolation with respect to the flux received from the Sun at the top of Earth’s atmosphere), transit depth, transit duration, and transit ingress time.

Transit signatures are also fitted in DV with a non-physical trapezoidal model [Ref] (as of SOC 9.3). The trapezoidal model fit parameters are epoch, transit depth, transit duration, and ratio of ingress time to transit duration. The orbital period is not fitted; the trapezoidal model fit employs the TCE period produced in TPS. The trapezoidal model fit is fast, and the model is employed later in DV as a fallback for the diagnostic tests that require a transit model in the event that a full transiting planet model fit result is unavailable for a given TCE.

Following model fitting and the multiple planet search, the next step in DV is to perform diagnostic tests on all planet candidates to aid in discrimination between true transiting planets and false positive detections. The diagnostic tests are performed sequentially and may be enabled or disabled on a test by test basis. Some of the diagnostic tests run very quickly and provide a large return on run-time investment. Other tests are time consuming and provide lower return on investment for the preponderance of planet candidates; these may be disabled when DV is run. The sequence in which the tests are (now) performed in DV is as follows: weak secondary test, rolling band diagnostic test, eclipsing binary discrimination tests, difference imaging and cen-

troid offset analysis, statistical bootstrap test, centroid motion test, and optical ghost diagnostic test. All of these tests were enabled for the final Q1–Q17 transit search (DR25).

DV data products are generated once the diagnostic tests have completed. These include a comprehensive DV Report in PDF format for each LC target with at least one TCE, a one-page DV Report Summary in PDF format for each TCE, a DV Time Series file in FITS format for each DV target including time series data relevant to the given target and associated TCEs [Ref], and a DV XML file including the aggregated DV results for all targets with TCEs in a given Pipeline run. The Time Series and XML files are not produced within DV proper, but by the Archive (AR) component of the *Kepler* Pipeline which is executed later. The DV data products are exported to the NASA Exoplanet Archive<sup>2</sup> at NExScI for access by the science community.

DV is executed on the NASA Advanced Supercomputing (NAS) Division Pleiades [Ref] computer cluster in a separate sub-task<sup>3</sup> for each LC target for which a TCE is generated in TPS. Pleiades is comprised of thousands of computing nodes in which multiple processing cores share common memory. Although there was a significant effort to reduce the DV memory footprint, this component is memory limited and does not utilize all available cores on each allocated processing node. For the Q1–Q17 DR25 processing, DV was run on Pleiades Ivy Bridge nodes with 20 processing cores and 64 GB of random access memory per node. DV was configured to allocate 6 GB per target and so utilized 10 of the 20 available cores on each node. In principle, all DV sub-tasks may be run in parallel, but in practice sub-tasks are queued and then processed as cluster resources become available.

All DV sub-tasks (one per target) running on Pleiades are subject to a maximum run-time limit (i.e., timeout). The planet search and model fitting process is allocated a configurable fraction of the total DV time limit (typically 0.8). The fitter and multiple planet search functions check periodically to determine whether or not the time allocation has been reached. If so, planet search and model fitting are halted to allow sufficient time for the remainder of DV to complete before the time limit is reached and the sub-task is killed. Furthermore, the time-consuming centroid motion (see Subsection 11.3.6) and optical ghost (see Subsection 11.3.7) diagnostic tests are also subject to self-timeout in that they are not run if insufficient remaining time would be available for DV report generation. Ultimately, there is no guarantee that DV will run to completion for all targets within the specified time limit.

The light curves of 198,707 targets were searched for transiting planet signatures in the Q1–Q17 DR25 TPS run; TCEs were generated for 17,230 of these (Twicken et al., 2016). The DV sub-task timeout was set to 45 hours, of which 36 hours were allocated to the fitter and multiple planet search. DV completed successfully for all 17,230 targets.<sup>4</sup> The median run time for all targets was 9.47 hours and the 90th percentile run time was 37.1 hours. The maximum run time was 44.8 hours, just below the 45 hour time limit at which point the sub-task would have been killed.

### 11.3 Diagnostic Tests

A suite of DV diagnostic tests is performed for each planet candidate identified in the Pipeline. These include TCEs identified in the initial TPS run for all LC targets and those subsequently identified in the multiple planet search with calls to TPS from DV. The diagnostic tests are described in this section. As stated earlier, the purpose of the tests is to produce metrics to aid in

<sup>2</sup><http://exoplanetarchive.ipac.caltech.edu>.

<sup>3</sup>Target stars in the *Kepler* FOV are assigned to “skygroups” representing the celestial regions which map to the respective CCD readout channels. The computational unit of work in DV includes targets in a given skygroup, so there is one Pipeline task for each of the 84 skygroups. Tasks are then subdivided into individual sub-tasks for each target.

<sup>4</sup>It should be noted that archive data products are not available for any target that does not successfully complete DV.

the discrimination between true transiting planets and false positive detections. Assessment and classification of the Pipeline candidates was described earlier in Subsection 11.1.1.

### 11.3.1 Weak Secondary Test

The purpose of the Pipeline transiting planet search is to identify signatures in *Kepler* target light curves that are representative of two-body Keplerian clocks. The TPS module was not designed to detect aperiodic signatures such as those associated with circumbinary transiting planets and planets with significant transit timing variations (TTVs). Nevertheless, the Pipeline has shown some sensitivity to TTV planets and detected many of them.

The most common false positive transiting planet detections are non-Keplerian in nature. The search for transiting planets by its nature must be extremely sensitive to small changes in the brightness of target stars to permit detection of Earth-size (and smaller) planets orbiting in the HZ of Solar-type stars. Non-Keplerian false positive detections are driven by a variety of sources including, but not limited to, electronic image artifacts [Ref], thermal variations and cycling [Ref], photometer pointing excursions [Ref], uncorrected or incompletely corrected sudden pixel sensitivity dropouts (SPSDs) [Ref], native stellar variability on transit time scales, and data gap edge effects.

False positive Keplerian detections may be ascribed to sources such as eclipsing binaries, background eclipsing binaries, planets transiting background stars, and contamination (e.g. saturation bleed, electronic crosstalk with neighboring readout channels or optical reflections) from bright eclipsing binaries or stars with transiting planets [Ref]. The most common false positive scenarios for Pipeline planet candidates that have been promoted to KOI status involve eclipsing binaries. Foreground or background eclipsing binaries may produce one or two TCEs depending upon eccentricity and the relative depths of the primary and secondary eclipses. The binary nature of a source is typically betrayed by a statistically significant match of the periods of the respective TCEs if two TCEs are generated.<sup>5</sup> If only one TCE is generated at a given period then the weak secondary test assesses the significance of the strongest secondary event at the same period and trial transit pulse duration. The diagnostic places a statistical constraint on the presence of secondary eclipses for each planet candidate identified in the Pipeline. The diagnostic also addresses the uniqueness, and hence the reliability, of the TCE itself [Ref].

The weak secondary diagnostic test produces a vector of multiple event detection statistics [Ref] as a function of phase for the period and trial transit pulse duration that produced the given TCE. For each phase value, the secondary Multiple Event Statistic (MES) represents a point estimate of the SNR of a sequence of secondary eclipses with the given period and trial transit pulse duration. The detection statistics are computed in the absence of the transits (or eclipses) that produced the TCE. The weak secondary algorithm is implemented in the TPS module where the transiting planet search is performed although the diagnostic test results are reported in DV and displayed in the data products.

The weak secondary algorithm is illustrated in Fig. ???. The period, epoch and trial transit pulse duration are determined when a TCE is generated in TPS. The transit signature that produced the TCE is removed by setting data gap indicators for the cadences associated with it; gap indicators for additional cadences preceding and following each of the transits are also set to provide a buffer against a trial transit pulse mismatch or relatively small TTVs. The light curve data gaps are then filled with the standard TPS gap filling algorithm [Ref]. A time-varying whitening filter is applied to the gap filled light curve to remove the statistical correlations in the

<sup>5</sup>The existence of two TCEs with matching periods on a given target does not imply that the source is necessarily an eclipsing binary; thermal and/or reflected light occultations of short period transiting planets may also produce transiting planet detections. We will see that DV models the secondary events to help ascertain whether or not they may be due to thermal or reflected light occultations.

time series, and the whitening filter is applied to the trial transit pulse for which the TCE was generated. Single Event Statistics (SES) are computed by correlating the whitened light curve with the whitened trial transit pulse in the same fashion that the transiting planet search is conducted [Ref]. The SES represent per cadence estimates of the single transit SNR for the given period and trial transit pulse duration.

The SES time series is folded at the period associated with the TCE and the detection statistics are combined [Ref] to form the secondary MES versus phase vector. The zero-point in phase corresponds to the epoch of the TCE. The maximum secondary MES is determined by the maximum value (over phase) of the secondary MES vector, and the minimum MES is determined by the minimum value of the secondary MES vector. In the absence of secondary eclipses, the multiple event detection statistics are expected to be zero mean and unit variance. The maximum secondary MES indicates the strength of a most significant secondary eclipse at the the period and trial transit pulse duration defined by the TCE. The minimum MES indicates the strength of the most significant positive-going signal at the period and trial transit pulse duration of the TCE.

The secondary MES vector is displayed versus phase (in units of days) in the DV Report with markers indicating the maximum and minimum secondary MES events. The following are tabulated in the DV Report: maximum secondary MES, phase of maximum secondary MES, minimum MES, phase of minimum MES, median absolute deviation of MES vector. The maximum secondary MES and associated phase are also indicated on the one-page DV Report Summary, which in addition displays the phase folded light curve associated with the given TCE with emphasis on the phase associated with the maximum secondary MES.

The weak secondary MES vector for KOI 140 based on Q1–Q17 DR25 data is shown in Fig. ???. The source of this false positive transiting planet detection is a background eclipsing binary that is offset by  $\sim 6$  arcsec from the target. The orbital period for the eclipsing binary is 19.978 days. For the TCE associated with the primary eclipses, the maximum MES reported by TPS is 128.6 for trial transit pulse duration = 9.0 hours. It is clear that there is a secondary eclipse present with phase = 9.222 days and MES =  $11.4\sigma$ . The minimum MES for this TCE =  $-3.6\sigma$ .

Given the large secondary MES and the Pipeline detection threshold of  $7.1\sigma$ , a second TCE would be expected for the secondary eclipses in the multiple planet search. Indeed, a second TCE was generated with maximum MES =  $11.9\sigma$  for period = 19.978 days and trial transit pulse duration = 10.5 hours. For this TCE, the maximum and minimum secondary MES were determined to be  $2.4\sigma$  and  $-3.0\sigma$  respectively.

The weak secondary MES vector for KOI 2887 based on Q1–Q17 DR25 data is displayed in Fig. ???. The source of this false positive transiting planet detection is offset by  $\sim 10$  arcsec from the target. The orbital period associated with the TCE is 1.569 days. A maximum secondary peak is visible for phase = 0.781 days with MES =  $5.1\sigma$ . The Pipeline would not be expected to generate a TCE for the secondary eclipses because the maximum secondary MES is less than the Pipeline detection threshold; indeed, a second TCE was not produced in this case. Nevertheless, the normalized phase ( $0.781/1.569 = 0.50$ ) associated with the maximum secondary event is highly indicative of a circular (i.e., eccentricity = 0) eclipsing binary. Inspection of the calibrated pixel time series data for this target leads to the same conclusion; the signature of an eclipsing binary is clearly visible in the pixels associated with the background source.

### 11.3.2 Rolling Band Diagnostic

A new diagnostic was introduced in the final revision of DV (SOC 9.3) to identify coincidences between transits and rolling band image artifacts [Caldwell Ref]. These temperature dependent image artifacts originate in focal plane electronics; the artifacts are particularly severe on a relatively small number of readout channels. The artifacts represent a major problem for the *Kepler Mission* because target stars rotate through these noisy channels for one observing quarter each

year; this leads to many false positive TCEs which when characterized appear to be small planets in long-period ( $\sim 1$  yr) orbits. Such orbits, of course, lie in the HZ of Sun-like stars. The rolling band diagnostic is described in this section.

### 11.3.3 Eclipsing Binary Discrimination Tests

We have shown that the weak secondary diagnostic test is capable of detecting the presence or constraining the significance of secondary eclipses associated with a given TCE. A set of statistical hypothesis tests are performed in DV to further aid in discriminating between transiting planets and eclipsing binaries. The binaries may be LC targets or background objects. The eclipsing binary discrimination tests are designed to flag the presence of an eclipsing binary if the system is circular and there is a single TCE, or regardless of eccentricity if there are separate TCEs for the primary and secondary eclipses.

After the transiting planet model has been fitted to all transits in the light curve associated with a given planet candidate, the model is fitted separately to the sequences of odd and even transits associated with the given candidate. A hypothesis test is performed to assess the statistical equality of the depth of the odd transits and the depth of the even transits. The odd and even transit depths for a true planet candidate would be expected to be consistent (although given quarterly spacecraft rolls, imperfect geometric placement of the CCDs, variations in detector performance across the focal plane, long time-scale focus variations, finite photometric apertures, and dynamic aperture crowding). The odd and even transit depths for a circular binary, however, would be expected to be inconsistent at some level due to differences in the characteristics of the stellar companions. We cannot guarantee that the difference is statistically significant, but we do perform the comparison and report the significance of the difference between the odd and even transit depths.

The difference in the epochs determined in the planet model fits to the sequences of odd and even transits is also compared statistically to one-half of the period associated with the fit to all transit events. An inconsistency in the timing of the sequences of odd and even transits would flag a slightly eccentric binary for which the transiting planet search has produced a single TCE. In reality, this eventuality almost never occurs. The  $\chi^2$  veto algorithms employed in TPS are very sensitive to such timing issues and almost categorically prevent the generation of a TCE when such issues are present (Seader et al., 2013). The odd/even epoch comparison diagnostic is still computed in DV because it was defined as a pre-launch requirement.

The final eclipsing binary discrimination metric computed in DV is a powerful one for flagging the presence of an eclipsing binary (foreground or background) when two (or more) TCEs are generated for a given LC target. In this case, the period determined in the planet model fit to all transits for a given candidate is compared statistically to (1) the period determined in the model fit for the candidate with the next shorter period (if one exists) and to (2) the period determined in the model fit for the candidate with the next longer period (if one exists).

A transiting planet detection is very likely to be a false positive if its period is statistically equivalent to that of another candidate associated with the same target. This would most commonly result from the generation of separate TCEs for the primary and secondary eclipses of a binary system. Multiple false positive TCEs may also result from significant stellar variability on the time scale of transits. Statistical equality of the periods of two planet candidates on a given target does not ensure that the candidates are not planetary, however. As discussed in Subsection 11.3.1, thermal and/or reflected light occultations for a short period planet may produce a TCE in addition to the one produced for the main transit events. Hence, the physical characteristics of highly compact systems must be examined closely in the cases where the shorter/longer period comparison diagnostics are statistically significant.

The eclipsing binary discrimination tests described above are implemented in DV as  $\chi^2$  hypothesis tests. Such a formulation supports the statistical comparison of multiple independent

measurements although only two are compared in each test. Consideration was also given to apply this formulation to assess of the consistency of (1) depths of all individual transits associated with a given planet candidate, and (2) transit depths determined separately from the quarterly data associated with each given planet candidate; such metrics were never implemented, however. The consistency check of  $N$  independent measurements of a parameter, denoted as  $x_i$ ,  $i = 1, 2, \dots, N$  with associated uncertainties  $\sigma_i$  is modeled as a statistical test with the null hypothesis that the  $x_i$  are drawn from  $N$  independent Gaussian distributions with the same mean value and standard deviations  $\sigma_i$ . The test statistic and significance level (i.e. p-value) are determined by

$$s = \frac{(x_1 - \bar{x})^2}{\sigma_1^2} + \frac{(x_2 - \bar{x})^2}{\sigma_2^2} + \dots + \frac{(x_N - \bar{x})^2}{\sigma_N^2} \quad (11.1)$$

and

$$p = \Pr(\chi_{N-1}^2 > s) \quad (11.2)$$

where  $\bar{x}$  is the weighted mean of the measurements  $x_i$  (with weights inversely proportional to  $\sigma_i^2$ ),  $\chi_{N-1}^2$  denotes a  $\chi^2$ -distribution with  $N-1$  degrees of freedom and  $\Pr()$  denotes “probability of”.

Acceptance of the null hypothesis for the equality of odd/even transit depths and odd/even transit epochs is consistent with a planetary classification for the transit source. Acceptance of the null hypothesis for equality in either of the shorter/longer period comparison tests, however, is not consistent with a planetary classification for the transit source. The standard in DV is to report diagnostic test significance such that values  $\sim 1$  are consistent with transiting planets (on target stars) and values  $\sim 0$  are inconsistent with transiting planets. Hence, the reported significance for the shorter/longer period comparison tests is reported as  $(1-p)$  with  $p$  as defined in Equation 11.2.<sup>6</sup>

It should be noted that for the purpose of the odd/even transit depth comparison test, the standard deviations  $\sigma_i$  are determined by the uncertainties in the respective transit depths as reported by DV. In the cases of the odd/even epoch test and the shorter/longer period comparison tests, however, the standard deviations  $\sigma_i$  are set equal to the transit durations derived from the fits to all transits for the respective planet candidates. The essence of the comparison in these cases is therefore to test that the transit timing and orbital periods agree to within the transit duration and not within the actual uncertainties in the fitted epochs and periods which are typically very small.

The phase folded odd and even transits are shown in Fig. ?? for KOI 6996 in the Q1–Q17 DR25 data set. The depth in each case is marked with a solid red line and the  $1\sigma$  uncertainties are marked with dashed red lines. The mismatch between the odd and even transit depths is clear. The difference reported for the odd/even transit depth comparison in this case was  $206\sigma$  ( $p = 0$ ). The source of this false positive transiting planet detection is a circular eclipsing binary.

The phase folded light curve is shown in Fig. ?? for KOI 140 in the Q1–Q17 DR25 data set. The blue event triangles mark the events that produced the first TCE for this target, and the red event triangles mark the events that produced the second TCE (at essentially the same period as the first). The significance of the shorter/longer period comparison test in this case was reported by DV to be  $(1-p) = 0.0005$ . The source of this false positive transiting planet detection is a background eclipsing binary.

<sup>6</sup>The significance of the eclipsing binary discrimination tests is commonly reported as a percentage rather than a fraction in the DV Report and one-page DV Report Summary, i.e.,  $100 \times p$  or  $100 \times (1-p)$  as applicable.

### 11.3.4 Difference Imaging and Centroid Offset Analysis

The intent of the weak secondary (Subsection 11.3.1) and eclipsing binary discrimination tests (Subsection 11.3.3) is to identify TCEs for which the source of a transit-like signature is likely to be an eclipsing binary (either foreground or background). DV also includes diagnostics designed to identify cases where the source of the transit (or eclipse) signature is likely to be a background star or stellar system. The goal of these diagnostics is to locate the source of the transit (or eclipse) signature; the offset between the source and target locations may be measured and its significance determined. The first of these diagnostics is difference imaging and centroid offset analysis which will be discussed in this section. The second diagnostic is the centroid motion test which will be discussed in Subsection 11.3.6. The utility of these diagnostics for identification of background false positives in *Kepler* data was documented by Bryson et al. (2013). Here we describe their implementation in the DV component of the *Kepler* Pipeline.

Difference imaging has proven to be a powerful diagnostic for identifying astrophysical false positive detections due to background sources (transits on background stars or background eclipsing binaries). The images are constructed from pixel data associated with each given target. The technique exploits spatial information contained in the pixel data and is capable of accurately identifying transit sources beyond the extent of the photometric apertures. This spatial information is not available in target flux and centroid time series. Difference images, PRF-based difference image centroids, and centroid offsets are computed on a quarterly basis (for each quarter in which transits are observed) for each TCE. The offsets are then robustly averaged over multiple quarters to improve the accuracy of the diagnostic. The centroid offset averaging is weighted by reciprocal uncertainties to emphasize offsets with relatively small uncertainties and deemphasize those with relatively large uncertainties.

For each planet candidate, mean in- and out-of-transit images are constructed by averaging the flux in and near each transit on a per pixel basis, and then by averaging over all transits for the given observing quarter. In- and out-of-transit cadences are identified from the transiting planet model that was fitted earlier to all transits in the target light curve. The difference image is produced by subtracting the mean in-transit flux value for each pixel from the mean out-of-transit flux. Uncertainties in the respective images are propagated from uncertainties in the calibrated pixel data by standard methods.

Transits are excluded from the respective difference images if the associated in- or out-of-transit cadences overlap (1) the transit of another planet candidate for the given target, (2) a known spacecraft anomaly (e.g., Earth-point for data downlink, safe mode, attitude tweak, and multiple-cadence loss of fine spacecraft pointing), (3) the start or end of the unit of work, or (4) cadences marked for exclusion by the Pipeline operator. The period following thermal events (return from Earth-point and safe mode) during which transits are excluded from difference image generation is a configurable parameter; typically this period was one day.

The pipeline may be configured, however, to prevent the exclusion of transits that overlap transits of another candidate associated with the same target if doing so would prevent the construction of a difference image in any given observing quarter; the rationale is that a possibly corrupted difference image is better than no difference image at all. Warnings are generated in such cases, but it is nevertheless true that such difference images may be difficult to interpret and are potentially misleading. DV was configured in this fashion for the Q1–Q17 DR25 run.

The photocenters of the out-of-transit and difference images are computed by fitting the appropriate Pixel Response Function (PRF) (Bryson et al., 2013) for the given channel and CCD coordinate position. The out-of-transit centroid locates the DV target, subject to aperture crowding (in extreme cases, the PRF-based centroiding algorithm locks onto a nearby star in the aperture mask that is brighter than the target). The difference image centroid precisely locates the source of the transit signature (which may or may not be the target) given sufficient SNR. The

quarterly offsets between out-of-transit and difference image centroids provide both absolute and statistical measures of the separation between target and transit source.

The offset is also computed per TCE and observing quarter between the difference image centroid and the target location specified by its celestial KIC coordinates. The offset from the KIC reference position is not subject to aperture crowding, but is subject to KIC errors and centroid bias. Robust mean centroid offsets are computed in right ascension and declination with associated uncertainties for the difference image centroids relative to the quarterly out-of-transit centroids and to the KIC coordinates of the given target. Difference image generation and centroid offset analysis will be described in detail in the following two subsections.

**11.3.4.1 Difference Image Generation** In-transit, out-of-transit, and difference images are generated for each DV target, planet candidate, and quarter as long as (1) the transiting planet model fit for the given planet candidate converged successfully or a trapezoidal model is available as fallback, and (2) there are one or more clean transits for the planet candidate in the given target table. A clean transit is one that occurred during a period when valid science data were collected, and one which is not excluded from the difference imaging process as described earlier. DV produces a so-called “direct” image displaying the mean flux per pixel over the duration of the quarter in the event that a difference image cannot be generated for a given planet candidate and observing quarter.

The difference image generation process is shown at a high level in Fig. ???. The iterations over quarters and planet candidates are illustrated. First, Pipeline data anomaly flags are parsed and anomalous cadences are defined. Transits are excluded from difference image generation if they overlap anomalous cadences in order to maintain the integrity of the results.

The in- and out-of-transit cadences are identified for all planet candidates over the duration of the unit of work. The model light curve is generated for each planet candidate and cadence based on parameter values of the transiting planet model fit to all transits (or trapezoidal model fit if planet model is unavailable). In-transit cadences are defined as those for which the transit depth exceeds a specified fraction (typically 0.75) of the maximum depth. Out-of-transit (i.e., “control”) cadences are defined as those sequences of cadences preceding and following each transit with durations equal to the transit duration derived from the model fit. A buffer (typically 3 cadences) is specified to isolate the control cadences from the transit events and preserve the integrity of the difference images in the event that the planet model fit is imperfect or there are moderate transit timing variations.

Transits corrupted by known data anomalies or by the transits of other planet candidates associated with the same target are excluded from the difference image generation process. This is to prevent compromising the quality and integrity of the difference image. Uncertainties in the resulting image values are larger than they otherwise would be if the corrupted transits were not excluded (because the averaging is performed over fewer transit events), but the image values are more accurate if the impacted transits are excluded. As stated earlier, it is possible to disable exclusion of overlapped transits if excluding the transits would prevent the generation of a difference image in any given quarter.

Identification of the cadences with known data anomalies has been discussed earlier. Transits are excluded from difference image generation if any cadence between the first and last out-of-transit control cadence (inclusively) is coincident with at least one of the (1) known data anomaly cadences, or (2) in-transit or associated buffer cadences for another planet candidate associated with the given target. Note that transits are not excluded from difference image generation if they are only coincident with the out-of-transit cadences of another planet candidate for the given target.

The process for formulating the mean in-transit, mean out-of-transit and difference images is illustrated in Fig. ???. It is performed in parallel for all pixels in the aperture mask associated with a given target. For each transit, the mean in-transit flux value is computed by averaging the cal-

ibrated pixel values (after removal of cosmic rays and background estimates) over the in-transit cadences and the mean out-of-transit flux value is computed by averaging the calibrated pixel values over the out-of-transit control cadences. Gapped (i.e., invalid or unknown) pixel values are ignored for the purpose of computing the mean fluxes and ultimately constructing the difference image. The total numbers of valid and gapped in- and out-of-transit cadences are included in the figure caption for each difference image displayed in the DV Report. Uncertainties in the mean in- and out-of-transit flux values for each transit are computed from uncertainties in the calibrated pixel values by standard methods under the assumption that the respective pixel values are not temporally correlated.

The overall mean in- and out-of-transit flux values are computed for each pixel by averaging the mean in- and out-of-transit levels associated with each of the transits over all of the transits in the given quarter. The difference image value for each pixel is then determined by subtracting the overall mean in-transit value from the overall mean out-of-transit value. Once again, uncertainties in the overall mean in- and out-of-transit flux values and in the difference value are computed by standard methods under the assumption that pixel values are not temporally correlated.

Fig. ?? illustrates the computation of the mean in- and out-of-transit values for one transit of Kepler-11e. Fifty cadences are displayed from the time series associated with the brightest pixel in the optimal aperture of Kepler-11 in Q4. The out-of-transit cadences used to determine the mean out-of-transit flux level for this transit are marked in green. The mean out-of-transit value is displayed as a horizontal green line. Uncertainties in the mean out-of-transit flux level are shown in black above and below the mean level. Control cadences both preceding and following the transit permit meaningful averages and differences to be computed without detrending (even for pixel time series with significant trends). The cadences used to determine the mean in-transit flux level are marked in red. The mean in-transit value is displayed as a horizontal red line with the associated uncertainties shown in black. The uncertainty in the mean out-of-transit level is less than that of the in-transit level because it is computed from a larger number of pixel values. The depth of this transit based on the out-of-transit level and difference shown here is 1350 ppm. The in- and out-of-transit cadences used to compute the mean levels for this transit were determined from the DV planet model fit to all transits in the corrected flux time series of this target.

The Q1–Q17 DR25 DV difference image diagnostic results for Kepler-11e (KIC 6541920) in Q5 are shown in Fig. ???. The overall mean out-of-transit values are displayed in the upper right panel as a function of the CCD coordinates<sup>7</sup> of the respective pixels in the target mask. The overall mean in-transit values are displayed in the lower left panel. The difference values are displayed in the upper left panel, and the difference SNR (value divided by uncertainty for each pixel) is displayed in the lower right. The pixels in the target mask are outlined in each panel with a solid white line, and the pixels in the optimal aperture are outlined with a dashed white line.

Kepler-11e is a confirmed planet; it is the largest of the six known transiting planets of Kepler-11. The scaling of the difference image values is nearly three orders of magnitude less than that of the mean out-of-transit values, but the visual character of the figures displayed in the two upper panels is essentially identical. The reference position for this target based on its KIC right ascension and declination is marked in all panels with a ‘x’; the centroid of the out-of-transit image is marked in the upper panels with a ‘+’; the centroid of the difference image is marked in the upper panels with a ‘Δ’. Centroiding of these images and centroid offset analysis will be discussed in Subsubsection 11.3.4.2. KIC objects near the target are marked with asterisks in all panels and identified by KIC ID and magnitude (Kp). The markers identifying target position

<sup>7</sup>The convention for numbering CCD rows and columns on the *Kepler* focal plane is that the pixel at the origin of the module output is row/column coordinate (0, 0). This is not a visible pixel, however; the origin of the photometric pixel region of each module output is row/column coordinate (20, 12) because the first 20 rows are masked and the leading 12 columns are virtual.

and difference image centroid are closely spaced; it is difficult to distinguish between target and transit source for this bona fide transiting planet.

The Q1–Q17 DR25 difference image diagnostic results for KOI 140 (KIC 5130369) in Q3 are shown in Fig. ???. KOI 140 is an astrophysical false positive (background eclipsing binary). The pixels with the largest flux differences for this planet candidate are clearly not coincident with the brightest pixels associated with the given target. In fact, the pixels with the largest flux differences do not even lie in the optimal aperture. The transit source as identified by the centroid of the difference image is clearly offset from the position of the target as indicated by both the KIC reference position and the out-of-transit centroid. In fact, the centroid of the difference image is nearly coincident with the position of KIC 5130380. This object is 2.5 magnitudes (10 times) fainter than the target and is almost certainly the source of the transit (i.e., eclipse) signature in the light curve of KOI 140.

We will briefly address the issue of when specific difference images can and cannot be trusted. In cases involving saturated transit sources (foreground or background), the difference images generally cannot be trusted; the transit signature is not visible in the pixels associated with the core of the transit source, but rather in the pixels at the ends of the bleeding column(s). In very low SNR cases, the difference images often cannot be trusted. In cases involving short time-scale stellar variability (time-scales comparable to the transit duration), the difference images cannot be trusted. A quality metric is computed in DV which appears to assess the respective difference images in a reliable fashion. The quality metric is computed by correlating the given difference image with the row/column pixel samples of the PRF when centered on the coordinates of the difference image centroid; sign is preserved so that quality metric values are in the range [-1, 1]. The value of the quality metric  $\sim 1$  if the shape of the difference image closely matches that of the PRF and the SNR is high; the value  $\sim 0$  when the difference image and PRF are uncorrelated. The quality metric for each quarterly difference image is compared against a configurable threshold (typically 0.7). The quality metric is considered reliable if it exceeds the threshold, otherwise it is not considered unreliable.

A summary quality metric is computed and reported for each TCE which is simply the fraction of quarterly quality metrics that exceed the specified threshold. The centroid offsets are considered reliable when a majority of the difference images from which they were computed are considered good. DV may also be configured to ignore the centroid offsets based on unreliable difference images (the so-called “quarter killer”). This functionality was not generally exercised other than for testing purposes. The issue remained concerning what to do in cases where most or all quarterly centroid offsets would be excluded from computation of the mean offset; it was not clear that such a result would be any more informative than the usual mean offset computation that does not account for difference image quality.

**11.3.4.2 Centroid Offset Analysis** Difference imaging is a powerful diagnostic for identifying false positive transiting planet detections due to background sources. This is accomplished by taking advantage of the spatial information inherent in the pixel time series to precisely locate the transit source in the photometric mask of the given target and determine the offset (in both absolute and statistical senses) between the transit source and the target itself. The target location is identified by two different methods. Each method has associated advantages and disadvantages which will be discussed later. Offsets are subsequently computed with respect to each of the target locations. In cases where the results are significantly different, the consumer of the DV products must decide which set of results (if any) to believe.

In the first case, the target CCD location is determined from its KIC right ascension and declination coordinates by evaluating so-called “motion polynomials,” and averaging over the valid in-transit cadences in the given quarter. Motion polynomials represent robust two-dimensional polynomial fits to the PRF-based centroids of 200 gold standard ( $K_p \sim 12$  and relatively uncrowded) targets on each module output; essentially these polynomials provide a cadence by

cadence mapping between the sky and the focal plane. The gold standard targets are the brightest for which the CCDs do not saturate, and therefore provide the highest fidelity centroids to determine the sky to focal plane mapping.

Evaluating the motion polynomials on the in-transit cadences and averaging the results allows the mean focal plane position of the target to be determined for the clean transits in the given target table.<sup>8</sup> The row and column coordinate estimates are assumed to be independent because the motion polynomials are separately computed in PA from row and column centroid coordinates and do not support the determination of row/column covariances.

In the second case, the target location on the CCD is determined for each quarter by computing the PRF-based centroid of the out-of-transit control image. The centroid aperture includes all pixels in the target mask. PRF-based centroiding is performed with a nonlinear fit that simultaneously solves for row/column translations and PRF scaling that best fit the pixel values in the given image. A row/column covariance matrix is produced for each centroid so that propagation of centroid uncertainties to later offset computations is not required to be performed under the assumption that row and column coordinates are independent. PRF-based out-of-transit centroids are transformed to sky coordinates by inverting motion polynomials and averaging over the valid in-transit cadences for the given quarter.

The location of the transit source is determined for each planet candidate and quarter by computing the PRF-based centroid of the respective difference image. As before, the centroid is transformed to sky coordinates with associated uncertainties. This PRF centroid represents the location of the transit source because the difference image is characterized by a star image centered on the transit source (assuming sufficient SNR).

Once the target and transit source locations have been computed, the centroid offsets are determined on both focal plane (in units of pixels) and sky (in units of arcsec). The process is illustrated in Fig.???. The magnitude of the offset is computed in each case as the quadrature sum of the two offset components. The uncertainty in the magnitude of each offset is computed by standard propagation of uncertainty methods. The centroid offsets are not computed if the difference image centroid cannot be successfully determined for a given planet candidate and observing quarter. Furthermore, the centroid offsets are only determined with respect to the KIC reference position if the difference image centroid is successfully computed, but the out-of-transit centroid is not.

The magnitude of the centroid offsets provides both absolute and statistical measures of the separation between the target and the transit source (which may of course be the target itself). The planet candidate is viable if the magnitude of the offset is statistically insignificant; it may still be the case that there is a background source (transiting or eclipsing) near the target, but it is not likely the case that there is a background source well separated from the target. The viability of the planet candidate must be called into question if the magnitude of the offset is significant; additional investigation is warranted in this situation.

As stated earlier, there are advantages and disadvantages associated with computing the centroid offsets with respect to each of the target locations described earlier. These must be understood to properly interpret the computed offsets. The out-of-transit image centroid is subject to crowding in the target mask whereas the difference image centroid is not. In a crowded centroid aperture it is therefore possible to obtain a significant offset with respect to the out-of-transit centroid even for a true transiting planet. The KIC reference position is not subject to crowding, but the centroid offset with respect to the KIC position is subject to KIC errors and biases in the PRF centroiding process itself. These biases tend to cancel when the offset is computed between PRF-based centroids for both out-of-transit and difference images, but the errors do not cancel when the offset computation involves only one PRF-based centroid.

<sup>8</sup>The target position on the focal plane is not static, but changes dynamically due to differential velocity aberration (DVA), temperature and focus variations, pointing variations, and commanded photometer pointing updates.

Centroid offsets are the principal tool employed in the vetting process to identify the presence of background sources. The offsets have been trusted on the order of 0.2 arcsec. For offsets less than 0.2 arcsec it is not generally believed that the presence of a background source can be confidently established. In order to prevent the offsets for high SNR candidates with small propagated centroid uncertainties from appearing to be significant when in fact they are not, a quadrature error term has been introduced into the computation of the mean centroid offset components and the magnitude of the mean offset. The value of this error term is a configurable Pipeline parameter. DV is typically run with a quadrature error term of  $0.2/3 = 0.0667$  arcsec; hence, the minimum  $3\sigma$  uncertainty in the magnitude of the mean offset is 0.2 arcsec, and no offset less than that is considered significant. The quadrature error term does not appreciably affect the vast majority of DV candidates for which the propagated uncertainties in the centroid offsets are much larger than 0.0667 arcsec. The quadrature error parameter may also be set to 0 in which case it is essentially removed from the offset analysis.

The difference image centroid offsets for Kepler-11e are shown in Fig. ???. The offsets of the quarterly difference image centroid relative to the out-of-transit image centroid are displayed in green in the left panel. The error bars indicate the  $1\sigma$  uncertainties in right ascension and declination for each offset. The offsets are also marked with the quarterly data set (“Qn”) from which they were computed. The quarterly offsets with respect to the KIC position of the target are displayed in the same fashion in the right panel. The robust weighted (by reciprocal uncertainties) mean offset over the 17 quarterly data sets is displayed in each panel in magenta with associated error bars. The  $3\sigma$  radius of confusion (i.e., three times the uncertainty in the magnitude of the mean offset) is displayed in each panel in blue. The target is located at the origin in each panel which lies comfortably within the respective radii of confusion. The transit source cannot be statistically differentiated from the target itself in each case. Kepler-11e is of course a confirmed transiting planet. The Q5 difference image for this target was shown in Fig. ???. Robust (to deemphasize outliers) averaging of multiple quarterly offsets improves the accuracy of the estimate of transit source location. The magnitude derived from the mean right ascension and declination offsets is  $0.1010 \pm 0.0819$  arcsec ( $1.23\sigma$ ) with respect to the out-of-transit centroid, and  $0.1365 \pm 0.0969$  arcsec ( $1.41\sigma$ ) with respect to the KIC position of the target.

The difference image centroid offsets for KOI 140 are shown in Fig. ???. The target is located at the origin in the offset reference frame which lies well outside the respective radii of confusion. KOI 140 is an astrophysical false positive (background eclipsing binary). The Q3 difference image for this target was shown in Fig. ???. The magnitude derived from the mean right ascension and declination offsets is  $5.801 \pm 0.073$  arcsec ( $79.4\sigma$ ) with respect to the out-of-transit centroid, and  $5.860 \pm 0.071$  arcsec ( $82.5\sigma$ ) with respect to the KIC position of the target. The robust mean offset in each panel hints strongly that the true source of the transit signature for this candidate is KIC 5130380 which is 2.5 magnitudes fainter than the given target. The centroid offsets are overlaid on UKIRT images in the lower panels of the figure; now it is clear that the source of the transit signature is not the target (identified with a red asterisk), but rather a nearby star.

### 11.3.5 Statistical Bootstrap

The purpose of the statistical bootstrap is to determine the false alarm probability associated with each TCE, i.e., the probability that a given TCE would have been generated with the same detection statistic due to noise alone in the absence of the transit signature. The false alarm probability is key to assessing TCE reliability [9.3 TCE paper ref]. The statistical bootstrap diagnostic employed in TPS and DV has been well documented [Refs]. The bootstrap as implemented in DV (SOC 9.3) is briefly discussed in this section.

### 11.3.6 Centroid Motion Test

It was shown in Subsection 11.3.4 that difference imaging may be utilized to identify the location of the transit source (which may be the target) associated with a given Pipeline TCE, and determine the offset of the transit source from the target in question. Centroid motion may alternatively be employed to locate the transit source and determine the offset of the transit source with respect to the target.

Flux-weighted centroids are computed for every target and cadence in the PA component of the Pipeline [Ref]. These identify the photocenter of the target within its centroid aperture in every *Kepler* image frame. Target centroid positions vary with time due to systematic effects discussed earlier. Centroid positions also vary as a result of changes in stellar brightness associated with transiting planets and eclipsing binaries. There is no centroid motion in principle for foreground transiting planets and eclipsing binaries when aperture crowding is negligible and the background is perfectly removed. In practice, however, all apertures are crowded to some degree and background removal is imperfect. Hence, centroids shift during transit or eclipse to a measurable extent in a majority of the cases. Whether or not the motion is statistically significant must be ascertained.

Centroids are computed in PA in the *Kepler* focal plane coordinate system, i.e., row and column index for a given CCD module and output. The flux-weighted centroid aperture includes the optimal photometric aperture plus a single halo ring [PA ref]. For the purpose of the centroid motion test, all centroids are first converted from focal plane to celestial coordinates (right ascension and declination) by inverting the motion polynomials computed on every cadence in PA (see Subsubsection 11.3.4.2).

The centroid motion test is performed for each planet candidate identified in the Pipeline. There are two aspects to the centroid motion test. We first seek to assess the degree of correlation between the centroid time series computed for the given target in PA and the model light curve derived from the DV planet model fit (or trapezoidal model fit if transiting planet model results are unavailable) to all transits for each associated TCE. It is unlikely that the transit signal is due to a background source if the degree of correlation is low. It is possible that the transit signal is due to a background source if the degree of correlation is significant. It is also possible that the target is the source of the transit signal, and that centroid motion is correlated with the model light curve as a result of aperture crowding or imperfect background removal. A centroid motion detection statistic is computed for each planet candidate that is distributed as a  $\chi^2$  random variable with two degrees of freedom; the significance of the statistic is also reported.

We also seek in the centroid motion test to determine the location of the transit source and in particular the offset between the transit source and the target itself (as determined by its KIC coordinates). The location of the transit source is expected to be consistent with the target location when the centroid motion detection statistic is insignificant. The location of the transit source may be inconsistent with the target in cases where centroid motion is significant. Flux-weighted centroids are a useful tool for differentiating between foreground and background transit sources, but it must be emphasized that the accuracy of the centroid test is dependent upon both target and transit source being well contained within the photometric aperture. As discussed in Subsection 11.3.4, source offsets determined by analysis of difference image centroid offsets are more reliable when the background source is beyond the photometric aperture.

A block diagram of the DV centroid motion test is shown in Fig. ???. The quarterly flux-weighted centroid time series are first converted cadence by cadence from CCD row and column coordinates to celestial coordinates by inverting PA motion polynomials that map between sky and focal plane. Systematic effects are then removed by cotrending the celestial centroid time series independently against spacecraft engineering data and motion proxies [Ref]. Centroid shifts due to brightness changes in the centroid aperture remain in the cotrended time series, but shifts common to the ensemble of targets on a given CCD are eliminated or at least highly

attenuated. The cotrended quarterly centroid time series are stitched together with compensation for level shifts and edge effects, and gaps in the time series are filled.

Transiting planet light curves derived from the transit model parameters for all candidates are jointly fitted (in amplitude only) to the right ascension and declination centroid time series. The process is performed iteratively in a whitened domain with the same machinery employed in the DV transit model fitter [Ref]. Detection statistics are computed separately for the right ascension ( $\alpha$ ) and declination ( $\delta$ ) centroid time series components. The detection statistics are defined for each planet candidate by

$$l_\phi = \frac{\tilde{b}_\phi \cdot \tilde{s}_\phi}{\sqrt{\tilde{s}_\phi \cdot \tilde{s}_\phi}}, \quad \text{for } \phi \in [\alpha, \delta] \quad (11.3)$$

where  $\tilde{b}_\phi$  is the whitened centroid time series component and  $\tilde{s}_\phi$  is the scaled whitened transit model for the given planet candidate. The detection statistic  $l_\phi$  should be significant if centroid motion in the direction of the associated celestial coordinate is correlated with the transit signature of the given planet and insignificant if there is no correlated motion in the direction of the coordinate. The transit signatures for all other planet candidates on the given target are first removed from the whitened centroid time series so that the detection statistics represent correlations only of the whitened transit model against the centroid time series signature associated with the given candidate.

The squares of the detection statistics are actually computed for each planet candidate in DV as the change in  $\chi^2$  for the respective fits [Ref]:

$$l_\phi^2 = \|\tilde{b}_\phi\|^2 - \|\tilde{b}_\phi - \tilde{s}_\phi\|^2, \quad \text{for } \phi \in [\alpha, \delta] \quad (11.4)$$

where

$$\|u\|^2 \equiv u \cdot u$$

For each planet candidate, the total centroid motion detection statistic is computed<sup>9</sup> as the sum of the squared statistics in each coordinate ( $\alpha, \delta$ ), that is

$$t = l_\alpha^2 + l_\delta^2 \quad (11.5)$$

The total motion detection statistic is distributed as a  $\chi^2$  random variable with two degrees of freedom. It is reported by DV for each planet candidate for which the transiting planet model fit is successful (and the iterative whitening and amplitude fitting process converges). The  $p$ -value for the total centroid motion detection statistic is given by

$$p = \Pr(\chi_2^2 > t) \quad (11.6)$$

The  $p$ -value represents the probability that a  $\chi^2$  statistic as large as  $t$  or larger would have been computed in the absence of correlated centroid motion due to random fluctuations in the centroids alone. This is reported as the significance for the test and follows the convention of the other statistical tests in DV. As stated earlier, it is likely that the transit source is the target itself if centroid motion is insignificant ( $p \sim 1$ ). Significant centroid motion ( $p \sim 0$ ) does not necessarily imply that the transit source is a background object, however. Centroid motion may be correlated with a transit signal on the target star because the photometric aperture is crowded or the background removal is imperfect.

The centroid shift during transit and the transit depth associated with a given planet candidate may be utilized to estimate the location of the transit source. For a fractional transit depth  $D$  that

<sup>9</sup>The definition of the centroid motion detection statistic was updated in SOC 9.3 to be  $t = l_\alpha^2 \cos^2(\delta_t) + l_\delta^2$  where  $\delta_t$  is the target declination. This definition weights motion equally in right ascension and declination.

is small compared to unity and an angular centroid shift  $\delta\phi$  during transit, the source offset  $\Delta\phi$  from the nominal out-of-transit centroid position may be estimated [Refs] by

$$\Delta\phi = -\delta\phi\left(\frac{1}{D} - 1\right) = -\delta\phi\left(\frac{1 - D}{D}\right) \quad (11.7)$$

The negative sign associated with  $\delta\phi$  in Eqn. 11.7 indicates that the centroid moves in the direction opposite that of the transit source when the source brightness decreases during transit. The uncertainty  $\sigma_\Delta$  in the source offset relative to the nominal out of transit centroid is given by standard propagation of uncertainties methodology as

$$\sigma_\Delta = \sqrt{\left(\frac{1 - D}{D}\right)^2 \sigma_\delta^2 + \left(\frac{\delta\phi}{D^2}\right)^2 \sigma_D^2} \quad (11.8)$$

where  $\sigma_\delta$  is the uncertainty associated with the centroid shift during transit and  $\sigma_D$  is the uncertainty in the transit depth.

The joint fit of the model light curves for the respective candidates to the centroid time series produces scale factors that identically represent the source offsets  $\Delta\phi$  with respect to the nominal out-of-transit centroid [Ref]. The peak centroid shifts  $\delta\phi$  and associated uncertainties are computed with Eqn. 11.7 and 11.8. The source offsets are subsequently added to the nominal out-of-transit centroid coordinates to estimate the absolute source location. Source offsets are then redefined with respect to the KIC position of the target by subtracting the KIC coordinates from the source location. Peak and source offsets in right ascension ultimately reported by DV are corrected by a cosine of target declination term to produce proper angular right ascension offsets (because one second in right ascension does not equal one arcsecond in angular offset except at the celestial equator). The magnitude of the peak centroid shift during transit is computed as the quadrature sum of the peak right ascension (corrected) and declination offsets, and the magnitude of the source offset is computed as the quadrature sum of the source right ascension (corrected) and declination offsets.

The following centroid test quantities are computed for each planet candidate and tabulated in the DV Report: total centroid motion detection statistic and significance, peak centroid shift in right ascension during transit, peak centroid shift in declination during transit, magnitude of peak centroid shift during transit, source offset from target location in right ascension, source offset from target location in declination, magnitude of source offset from target location, absolute source right ascension coordinate, and absolute source declination coordinate. Uncertainties in all quantities but motion detection statistic are also computed and tabulated in the DV report. Peak centroid offsets during transit, source offsets from target star, and all associated uncertainties are reported in units of arcsec.

Centroid motion test results for KOI 140 in the Q1–Q17 DR25 data set are shown in Fig. ???. Detrended and phase folded flux values are shown in the upper panel, detrended and phase folded right ascension centroid shifts are shown in the middle panel, and detrended and phase folded declination centroid shifts are shown in the bottom panel. The scaled transit model is overlaid on both the right ascension and declination centroid figures. Although the amplitude fitting is performed in the whitened domain, the diagnostic figures are displayed in the unwhitened domain. The magnitude of the peak centroid shift during transit is 13.9 milliarcsec ( $142\sigma$ ). The total centroid motion detection statistic is very large (32,100) and the significance of the statistic is essentially 0. Centroid motion is incontrovertible.

The source offset is determined to be 19.8 arcsec ( $146\sigma$ ). This overestimates the true source offset for this false positive KOI where the background eclipsing binary source is known to be 5.8 arcsec from the target. The discrepancy is due to the fact that the background source lies on the boundary or outside the photometric aperture of the target in most quarters. The true transit depth is underestimated which leads to an overestimate of the source offset. This is discussed more fully in [Bryson ref].

Centroid motion test results for Kepler-62f in the Q1–Q17 DR25 data set are shown in Fig. ???. The transit signature is clearly visible in the detrended and phase folded flux displayed in the top panel, but there is little discernible centroid offset in either right ascension or declination. The magnitude of the peak centroid shift is determined to be 0.29 milliarcsec ( $0.87\sigma$ ). The centroid motion detection statistic is 2.57 for which the significance is 0.28 (not statistically significant). The magnitude of the source offset from the KIC position of the target is estimated to be 1.08 arcsec ( $1.50\sigma$ ). The fitted transit depth for this confirmed HZ super-Earth is  $470 \pm 31$  ppm.

### 11.3.7 Optical Ghost Diagnostic Test

A new diagnostic test was introduced in the final revision of DV (SOC 9.3) to identify planet candidates for which a TCE was likely generated due to optical ghosts (or well distributed contamination). The test involves correlating flux time series derived from aperture core and halo pixels against the transit model light curve for the given TCE. The aperture core flux time series should be more highly correlated with the planet model if the target is the transit source. The halo flux time series may be more highly correlated with the planet model if the transit source is an optical ghost or distributed contamination. The optical ghost diagnostic test is discussed in this section.

## 11.4 KOI Matching

All TCEs identified in the *Kepler* Pipeline transit search and fitted in DV may be optionally matched against the ephemerides of known KOIs as of the time that DV is executed. The ephemerides are obtained by a Pipeline operator from the cumulative KOI table at the Exoplanet Archive at NExSci. Ephemeris matches at the target (e.g., KOI 157) and planet (e.g., KOI 157.01) levels are reported in the DV archive products for the benefit of the *Kepler* Project and user community. KOI matching was enabled in DV for the Q1–Q17 DR24 and DR25 runs. The algorithm implemented in DV for matching ephemerides of known KOIs and Pipeline TCEs is discussed in this section. A different matching algorithm is employed by TCERT for federating Pipeline TCEs with existing KOIs [Ref].

## 11.5 Archive Products

DV produces archive products for export to the science community that summarize the information that is provided to DV and encapsulate the results of the planet model fits and diagnostic tests within DV. We reiterate that the design goal of DV was not to determine the likelihood that a particular TCE is a legitimate transiting planet; rather the design goal of DV was to characterize each TCE and perform a uniform set of diagnostic tests to enable consumers of DV products to vet the TCEs and make assessments or classifications. The DV products can only be briefly summarized here. Space does not permit a complete description of all aspects of these products. It should also be noted that the DV products evolved with each release of the *Kepler* Pipeline code base. The descriptions provided apply to SOC 9.3 which was employed for the Q1–Q17 DR25 transit search.

DV generates four types of archive products. Comprehensive DV Reports are produced in PDF format for each target with at least one TCE; the Report is summarized in Subsection 11.5.1. One-page DV Report Summaries are produced in PDF format for each TCE; the Report Summary is summarized in Subsection 11.5.2. DV Time Series files are produced in FITS format for each target with at least one TCE; this product is summarized in Section 11.5.3. Reports, Report Summaries, and Time Series files are exported to the NASA Exoplanet Archive (see Section 11.2).

where they are available to the user community and general public. Finally, DV produces a single XML file containing the tabulated results for all targets in a given DV run. This file is used to populate tables at the Exoplanet Archive. Although it is a text file, it is not considered to be human readable and will not be discussed further in this publication.

### 11.5.1 DV Report

DV produces a comprehensive Report in PDF format for each target with at least one TCE in a given run. The Reports are automatically generated with LaTeX and delivered to the NASA Exoplanet Archive at NExScI where they are accessible by the user community and general public. The Report is organized into logical sections; these will be summarized below. The PDF files include tabs for sections and sub-sections to allow users to quickly locate the specific DV results that they seek. The Report evolved significantly over the course of the *Kepler Mission*.

**11.5.1.1 Summary** Following a cover page and table of contents, the report begins with a Summary. This may be considered an executive summary; if a user only wants the basic DV results for a given target it may not be necessary to delve any further than this.

The summary includes tables with target properties, data acquisition characteristics, and planet candidate properties. The target properties represent the stellar parameters (and associated uncertainties) that are provided to DV: magnitude, location, radius, effective temperature, surface gravity, and metallicity. A provenance string is included for each stellar parameter to inform users about the source of the information. Keys to the provenance strings are published separately.

The data acquisition characteristics table includes one entry for each quarter in which the given target was observed. For each quarter, the table specifies the quarter, the CCD module output, the crowding metric and flux fraction in aperture utilized to correct the light curve in PDC [Ref], and the limb darkening coefficients. DV was designed to accommodate quarter (and hence module output) specific limb darkening coefficients but this functionality was never deemed sufficiently necessary to implement. Hence, the set of limb darkening coefficients does not change on a quarterly basis.

The planet candidate characteristics table includes one entry for each TCE associated with the given target. The table specifies period, epoch, semi-major axis, planet radius, and equilibrium temperature for each DV candidate along with a flag to indicate whether DV suspected the candidate to be an eclipsing binary (based on transit depth alone) and therefore omitted the planet model fits which do not implement an eclipsing binary model.

DV was updated in SOC 9.2 to include KOI numbers and Kepler names (for confirmed planets) where applicable in the target and planet properties tables. Matches at the target level are determined by Kepler ID; matches at the planet level are determined by correlating KOI and DV model fit ephemerides as described in Section 11.4. We reiterate that the KOI and Kepler name information displayed in the DV archive products pertain to known KOIs at the time that DV was executed; new KOIs identified from the TPS/DV results of a given run will not be identified as such in the archive products produced for that particular run. We also note that the matching process at the planet level is not perfect. The Report includes a list of planet-level KOIs that could not be matched successfully against the Pipeline results for the given target.

**11.5.1.2 UKIRT Images** The celestial context in the vicinity of the DV target can be invaluable for digesting and interpreting the DV diagnostics that attempt to establish the position of the transit source with respect to the target. To that end, we have downloaded images from the UKIRT survey [Ref] for nearly every target that has appeared on a *Kepler* target list. For each target, the image displayed in the DV Report covers a region approximately one arcmin square. Difference image centroid offsets are also displayed on UKIRT images for the associated target stars. Right ascension and declination grid lines are overlaid on the UKIRT images. We were unable to obtain images for all Kepler targets due to lack of coverage in the survey data.

**11.5.1.3 Flux Time Series** The quarter-stitched PDC (i.e., systematic error corrected) light curve is displayed with markers to indicate the transit times of the various candidates associated with the given target. This is the light curve that is first subjected to the transiting planet search. The light curve is segmented by quarter, and each quarter is displayed separately with vertical offsets for clarity. As part of the quarter-stitching process, the quarterly segments are normalized and strong harmonic content is removed. Gaps are filled in the quarter-stitching process, but gap filled data are not displayed in this section. Gaps for monthly data downlinks and spacecraft safe mode are clearly visible in these figures. These figures are particularly valuable diagnostic tools for planet candidates based on relatively few transits. The detection is suspect if the transit markers in such cases overlay uncorrected or partially corrected SPSDs or spacecraft attitude adjustments, fall on the boundaries of data gaps, or occur during particularly noisy data segments.

The quarterly PA (i.e., SAP) light curves are also shown in this section. These are also valuable diagnostic tools because the data are displayed prior to systematic error correction in PDC and quarter-stitching in TPS/DV. Gross discrepancies between the PA and quarter-stitched PDC light curves imply that the post-PA processing has been off nominal for the given target. For example, short period transits may be inadvertently eliminated in some or all quarters in the harmonics identification and removal function within the quarter-stitching process. It is a red flag if the transits are clearly visible in the PA light curves, but are not present in the quarter-stitched PDC light curves; the light curve in question has been corrupted in a well intentioned attempt to improve the sensitivity to the most valued planet candidates in the *Kepler Mission*.

**11.5.1.4 Dashboards** There is one dashboard figure for each planet candidate associated with the given target. The dashboards summarize the model fit and many DV diagnostic test results. Each region on the dashboard figure is color coded; the caption on the dashboard fully explains the coding. In general, nominal results are displayed in green, borderline results in yellow, and results that may call the planetary nature of any TCE into question are displayed in red. The regions are colored blue when results are unavailable.

The dashboard provides a means to view the DV results at a glance and focus quickly on issues pertaining to any given candidate. It must be emphasized, however, that if a region is colored red the candidate may still be planetary in nature. We discussed in Subsection 11.3.3 that short period planets with detectable occultations may trigger the eclipsing binary discrimination test for equal periods. We also discussed in Subsection 11.3.6 that there may be significant centroid motion during transit for targets with transiting planets in crowded fields. In neither of these cases does red coloring invalidate the planetary nature of the candidate.

**11.5.1.5 Centroid Cloud Plot** The change in flux is displayed versus change in right ascension (blue) and declination (red) centroid coordinates. The flux and respective centroid time series are unwhitened and median detrended. In-transit centroid motion manifests itself as a deviation from the vertical below the jitter cloud. The centroid cloud plot is a coarse representation of the centroid motion statistic and peak in-transit centroid offsets discussed in Subsection 11.3.6 as part of the centroid motion diagnostic test. If correlated centroid motion is present in the centroid cloud plot then its presence is incontrovertible. Significant centroid motion may still be present, however, if correlated centroid motion is not visible in the centroid cloud plot.

**11.5.1.6 Image Artifacts** The rolling band contamination diagnostic (see Subsection 11.3.2) results are displayed in a table for each DV planet candidate. The table indicates the number of transits (and fraction of total) that are coincident with rolling band image artifacts at each of the defined severity levels. As discussed earlier, the severity levels range from 0 (low) to 4 (high). The reliability of a TCE is questionable if a significant fraction of the total number of transits are coincident with severity levels  $> 0$ , particularly for long-period TCEs with relatively few transits. Individual transits with non-zero severity levels are highlighted in a diagnostic light curve panel on the one-page DV Report Summary.

**11.5.1.7 Pixel Level Diagnostics** The pixel level diagnostic test results are displayed separately for each planet candidate associated with the given target. The difference image summary quality metrics are presented in a table; the table includes the correlation threshold that defines the cutoff between good and bad quality difference images. The difference image centroid offsets discussed in Subsubsection 11.3.4.2 are displayed in both graphical and tabular form. Offsets are displayed with respect to the out-of-transit centroid and with respect to the KIC position of the target. Robust mean results are displayed for all TCEs. The value of the error term that is added in quadrature to the robust mean and bootstrap PRF fit offsets is included in the figure captions. The offsets are also overlaid on the UKIRT image associated with the given target if such an image is available.

The difference images are subsequently displayed quarter by quarter. The caption for each difference image includes the value of the quality metric for the given quarter. The caption also indicates the number of transits and valid cadences that were used to compute the difference image for the given quarter, and the number of in- and out-of-transit cadence gaps. Quarterly PRF centroid results and centroid offsets are tabulated on the focal plane (in units of pixels) and on the sky (in units of arcsec). Nearby KIC and UKIRT catalog objects are marked on the respective difference images, as are the image centroids and target KIC position.

**11.5.1.8 Phased Light Curves** Full phase-folded light curves are displayed in both unwhitened and whitened domains for each of the planet candidates associated with the given target. Colored event triangles below each figure mark the phase of the transits associated with all of the candidates on the target. The phased light curves are particularly useful in multiple TCE systems to study the phase relationships between the candidates. This applies to multiple planet systems which may have resonant relationships between candidates and to binary systems where primary and secondary eclipses have a common period but different phase. The phased light curves can also highlight false detections in multiple planet systems due to image artifacts where multiple planet candidates may “transit” on the same module outputs with long periods that are not identical but similar; in these cases the event triangles for the false detections share a common region of phase space and appear in clusters.

Beginning in SOC 9.2, median detrending is applied to the unwhitened data prior to phase folding. In earlier releases, the unwhitened data were not detrended prior to phase folding.

In SOC 9.3, phase-folded light curves by quarter, by observing season<sup>10</sup>, and by year are displayed for each planet candidate. These phase-folded light curves are based on median detrended, unwhitened data.

**11.5.1.9 Planet Candidate Results** The bulk of the planet model fit and diagnostic test results are presented in a separate section of the DV Report dedicated to each planet candidate. Each section begins with tables containing the TCE parameters for the given candidate and the results of the model fit to all transits. Fit results include parameter values and associated uncertainties. The quarter-stitched PDC light curve for the given candidate is displayed in quarterly segments with markers highlighting the transit events. This differs from the quarter-stitched PDC light curve described in Subsubsection 11.5.1.3 in that transits for all DV candidates prior to the given one have been removed. Essentially, the light curve displayed here is the one in which the transiting planet detection was made for the given candidate in TPS.

Diagnostic figures illustrating the phase-folded flux time series data in the unwhitened and whitened domains are presented. The whitened transit model is overlaid on the phase folded data in the whitened domain. Colored markers differentiate between the data points that were included and emphasized in the robust model fit and those that were deemphasized or otherwise

<sup>10</sup>Kepler observing seasons are denoted by S0, S1, S2, S3. Each season corresponds to a specific photometer roll orientation. As discussed earlier, the photometer was rolled by 90° every quarter in order to main illumination of the solar panels.

ignored. Reduced parameter fit results are displayed graphically and in tabular form as a function of impact parameter. The quality of the fit results are often only weakly dependent on impact parameter so the reduced parameter fits may represent a family of equally valid results for the given planet candidate. This information is useful to the community because it clarifies that in many cases the planet characteristics are not uniquely determined by planet model fitting.

Weak secondary test results are displayed both graphically and in tabular form. The table includes the detection statistic for the given TCE as well as the maximum and minimum secondary detection statistics (and associated phases). This test was described in Subsection 11.3.1. The centroid motion test results (see Subsection 11.3.6), eclipsing binary discrimination test results (see Subsection 11.3.3), statistical bootstrap test results (see Subsection 11.3.5), and optical ghost diagnostic test results (see Subsection 11.3.7) are displayed in separate tables. The centroid motion results are derived from flux-weighted centroids that are computed in PA for all targets and cadences. Finally, a series of diagnostic figures are presented illustrating flux-weighted centroid motion for the given candidate. The detrended phase-folded flux and centroid time series are shown first, followed by quarterly figures that mark the transit times on the respective flux and centroid time series.

**11.5.1.10 Appendices** The Appendices to the DV Report contain valuable diagnostic information in spite of the fact that the figures and tables were not deemed sufficiently important to be displayed in the main body of the document. The robust weights for the planet model fit to all transits for each candidate associated with the given target are displayed as a time series and also with folded phase. Issues with the robust planet model fit may be highlighted by irregularities in the figures. Histograms of fit residuals for constraint points and all valid data points are also displayed with Gaussian overlays.

Results of the model fits to the sequences of odd and even transits are displayed for each candidate in tabular form. These support the eclipsing binary discrimination tests discussed in Subsection 11.3.3. Of particular interest are the transit depths and associated uncertainties for the odd and even transit fits. The difference in the fitted depths for the odd and even transits divided by the uncertainty in the difference essentially determines the significance of the odd/even depth comparison test.

Diagnostic figures illustrating the phase folded flux time series data in both unwhitened and whitened domains are presented for the odd and even transit model fits for each candidate. As before, the whitened transit model is overlaid on the phase folded data in the whitened domain. Colored markers differentiate between the data points that were emphasized in the respective robust model fits and those that were deemphasized or otherwise ignored.

**11.5.1.11 Alerts** Alerts are generated at run time in DV (and the other Pipeline components) to flag off-nominal conditions. Pipeline alerts are categorized as either Warnings or Errors. The alerts issued by DV largely flag Warning conditions only. An alert contains a time stamp, severity (i.e., “warning” or “error”) condition, and message string. DV alert message strings include the KIC ID of the target, the index of the planet candidate where applicable, and the name of the DV sub-component in which the alert was raised. The alerts were originally implemented to support the operation and maintenance of the Pipeline, but it was also decided to include the alerts in the DV Report as a service to the user community.

## 11.5.2 DV Report Summary

DV produces a one-page Report Summary in PDF format for each TCE identified in the transit search. The Report Summary includes useful diagnostic figures and tabulated model fit and diagnostic test results; it has proven to be extremely beneficial for assessing the character of DV planet candidates. A useful companion for the SOC 9.3 version of the Report Summary produced

for the Q1–Q17 DR25 TCEs was produced by the *Kepler* Project, and is hosted at the Exoplanet Archive.<sup>11</sup>

### 11.5.3 DV Time Series

A DV Time Series file in FITS format is generated for each LC target with at least one TCE by the AR component of the *Kepler* Pipeline after all DV tasks are completed. The file includes time series data relevant to the processing of each given target in TPS/DV and all associated TCEs. The DV Time Series file was extensively redesigned in SOC 9.3. The SOC 9.3 Time Series file content applicable to the Q1–Q17 DR25 transit search was documented by [Thompson Ref]; this internal *Kepler* document is hosted at the Exoplanet Archive.<sup>12</sup>

## 11.6 Conclusion

The paper is summarized and concluded in this section.

## Bibliography

- Akeson, R. L., Chen, X., Ciardi, D., et al., 2013. “The NASA Exoplanet Archive: Data and Tools for Exoplanet Research,” PASP, 125, 989
- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A., 2011. “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” AJ, 142, 112
- Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al., 2013. “Identification of Background False Positives from Kepler Data,” PASP, 125, 889
- Claret, A., & Bloemen, S., 2011. “Gravity and Limb-Darkening Coefficients for the *Kepler*, CoRoT, Spitzer, uvby, UVBVRJHK, and Sloan photometric systems,” Astronomy & Astrophysics, 529, A75, available online at <http://dx.doi.org/10.1051/0004-6361/201116451>
- Huber, D., Silva Aguirre, V., Matthews, J. M., et al., 2014. “Revised Stellar Properties of Kepler Targets for the Quarter 1–16 Transit Detection Run,” ApJS, 211, 2
- Jenkins, J. M., 2002. “The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets,” ApJ, 575, 493
- Mandel, K., & Agol, E., 2002. “Analytic Light Curves for Planetary Transit Searches,” ApJL, 580, L171
- Seader, S., Tenenbaum, P., Jenkins, J. M., & Burke, C. J., 2013. “ $\chi^2$  Discriminators for Transiting Planet Detection in Kepler Data,” ApJS, 206, 25
- Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016. Kepler Archive Manual (KDMC-10008-006) (Moffett Field, CA: NASA Ames Research Center)
- Twicken, J. D., Jenkins, J. M., Seader, S. E., et al., 2016. “Detection of Potential Transit Signals in 17 Quarters of Kepler Data: Results of the Final Kepler Mission Transiting Planet Search (DR25),” AJ, 152, 158

<sup>11</sup><http://exoplanetarchive.ipac.caltech.edu/docs/DVOnePageSummaryPageCompanion-dr25-V7.html>.

<sup>12</sup><http://exoplanetarchive.ipac.caltech.edu/docs/DVTimeSeries-Description.pdf>

- Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010. "Data Validation in the Kepler Science Operations Center Pipeline," in Proc. SPIE, Vol. 7740, 42W



## CHAPTER 12

---

# DATA VALIDATION II – TRANSIT MODEL FITTING AND MULTIPLE PLANET SEARCH

JIE LI<sup>1</sup>, PETER TENENBAUM<sup>1</sup>, JOSEPH D. TWICKEN<sup>1</sup>, CHRISTOPHER J. BURKE<sup>1</sup>, JON M. JENKINS<sup>2</sup>, ELISA V. QUINTANA<sup>1</sup>, JASON F. ROWE<sup>1</sup>, SHAWN E. SEADER<sup>1</sup>,

<sup>1</sup>The SETI Institute, Mountain View, CA 94043, <sup>2</sup>NASA Ames Research Center, Moffett Field, CA 94035

**Abstract.** This chapter discusses the transit model fitting algorithms and performance of the *Kepler* Science Data Processing Pipeline. Threshold Crossing Events (TCEs), which are transit candidate events, are generated by the Transiting Planet Search (TPS) component of the pipeline and subsequently processed in the Data Validation (DV) component. The transit model is used in DV to fit TCEs in order to derive parameters that are used in various diagnostic tests to disposition planetary candidates. Transit model fitting performance of the latest release (9.3, January 2016) of the pipeline is demonstrated with the results of the processing of 4 years (17 quarters) of flight data. The transit model fitting results of the TCEs are accessible from the NASA Exoplanet Archive.

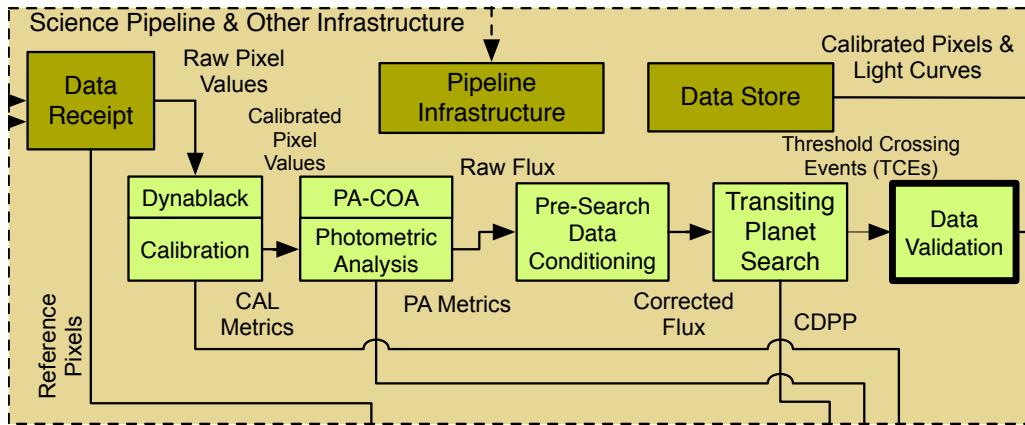
**Keywords:** *Kepler*, light curve, planets and satellites: modeling.

**Facilities:** The *Kepler* Telescope

### 12.1 Introduction

This chapter discusses transit model fitting algorithms and performance that are part of the Data Validation (DV) module of the *Kepler* Science Data Processing Pipeline. The overall architecture of DV is described in Chapter 11, which also details the various diagnostic tests performed on each TCE. Figure 12.1 shows the DV module in the context of the *Kepler* Pipeline.

The transit model fitting is designed for the following two main tasks: 1) The fitted parameters of the transit model and the corresponding light curve generated from the model are used in the diagnostic tests in DV to disposition planetary candidates; 2) When the TPS module is called, only one TCE with the maximum multiple event detection statistic (MES) is generated. To search for multiple-planetary candidates, an iterative process of transit model fitting and multiple-planet search is implemented in DV. For each target star, the transit model parameters are fitted to each TCE generated by TPS, the signature of known TCEs is removed from the light curve and the flux time series, and then the residual goes through TPS again to search for additional TCEs. This iteration will only terminate once no new TCEs are identified or a preset upper limit is reached (set to 10 for the SOC 9.3 run producing the DR 25 TCEs).



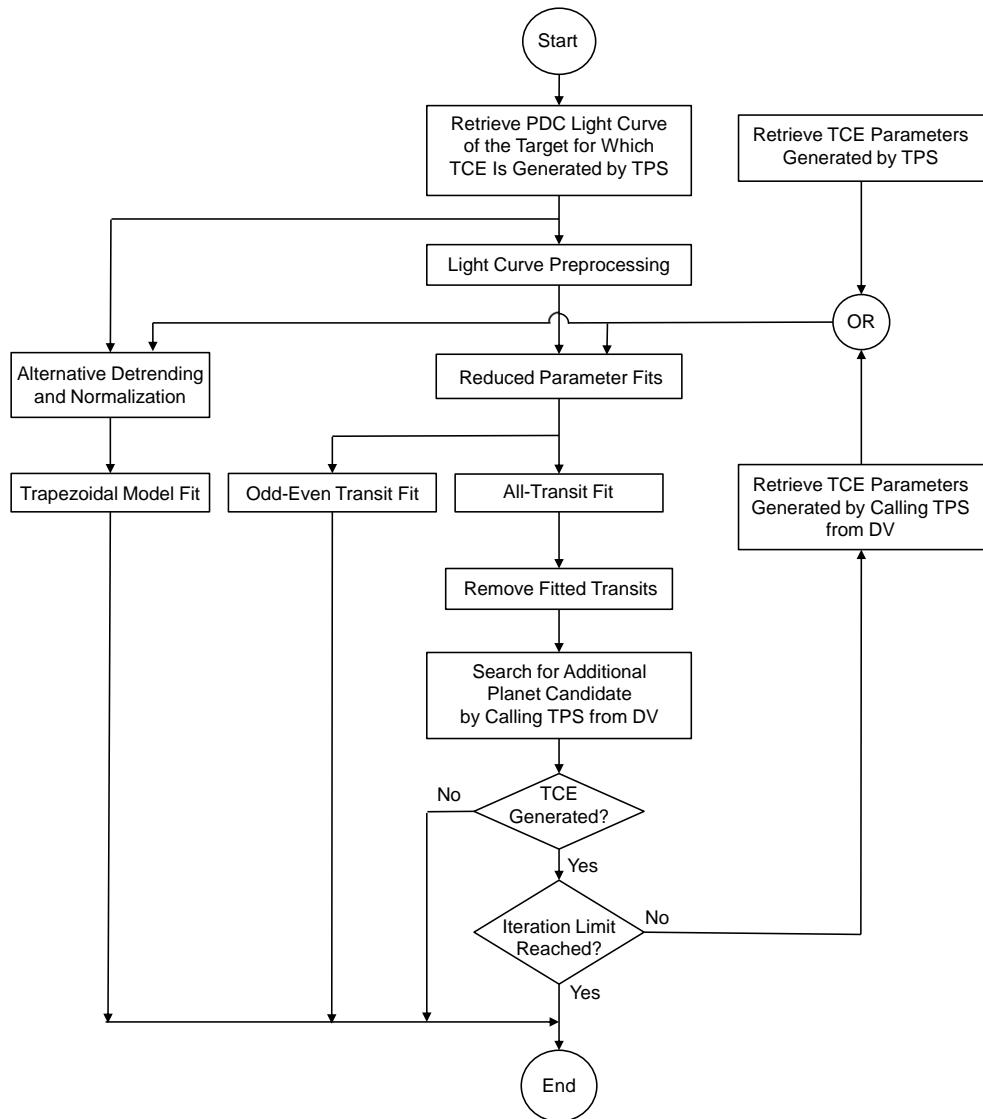
**Figure 12.1** Data Validation (DV) in the context of the architecture of the *Kepler* Data Processing Pipeline. As a part of the DV module, the transit model fitting fits TCEs generated by TPS to derive parameters used in various diagnostic tests of DV.

The transit model fitting results, such as the fitted parameters and uncertainties, derived parameters and uncertainties, goodness of fit metrics, and the diagnostic plots, are included in the DV report or DV one-page summary report, both accessible by the science community at the NASA Exoplanet Archive<sup>1</sup>.

Figure 12.2 shows the flowchart of the iterative process of transit model fitting and multiple-planet search. When a TCE is generated by the TPS module, the corresponding systematic error-corrected light curve of the target star, generated by the PDC module, is furnished to DV along with the transit parameters associated with the TCE, including the transit epoch and orbital period, transit duration, and MES of the TCE. After several preprocessing procedures described in Section 12.2, the light curve of the target star goes through a series of transit model fitting algorithms, which include the reduced parameter fits, all-transit fit, odd-even transit fit and the trapezoidal model fit.

As shown in Figure 12.2, the light curve after preprocessing is first applied to a set of reduced parameter fits, in which the impact parameter is set to fixed values of 0.1, 0.3, 0.5, 0.7 and 0.9, and only the parameters of transit epoch time, planet orbital period, ratio of planet radius to star radius and ratio of semi-major axis to star radius are fitted to a geometric transit model, which is defined in Section 12.3. The initial values of the fitted parameters of the reduced parameter fits are determined from the TCE parameters. The reduced parameter fits resolve the degenerated problem of fitting the impact parameter. After the completion of the reduced parameter fits, all-transit fit and odd-even transit fit follow, in which the fitting algorithms are applied to all transits, odd transits and even transits, respectively. The all-transit fit and odd-even transit fit are both initialized with the fitted parameters of the reduced parameter fit with the minimum  $\chi^2$  metric. The output of the all transit fit is used in several diagnostic tests of DV and the assessment of the planet candidacy, and the output of the odd-even transit fit is used in a specific DV diagnostic test to identify false positives generated by an eclipsing binary target or a target with an eclipsing binary in the background. Section 12.4 describes how a synthetic light curve is generated from the fitted parameters of the geometric transit model and Section 12.5 describes the algorithms to fit the light curves with the geometric transit model in detail. In addition to the fitting algorithms with the geometric transit model, a fitting algorithm with the trapezoidal model is described in

<sup>1</sup>The most recent TCE table can be viewed at <http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-TblView?app=ExoTbls&config=tce>.



**Figure 12.2** Flowchart of iterative process of transit model fitting and multiple-planet search.

Section 12.6. As shown in Figure 12.2, an alternative detrending and normalization algorithm is applied to the PDC light curve prior to the trapezoidal model fit. The output of the trapezoidal model fit is used in the diagnostic tests of DV when the fit with the geometric transit model fails or when the fit is not performed, e.g. for suspected eclipsing binaries.

After the completion of the transit model fitting algorithms, the signature of the known TCE, whose location is determined from the fitted parameters of the all-transit fit, is removed from the light curve, and the residual light curve is subject to a search for additional planets by calling TPS in the DV module. If an additional TCE is generated, the residual light curve goes through the transit model fitting algorithms discussed above once again. The multiple-planet search is discussed in Section 12.7. As shown in Figure 12.2, the iterative process of the transit model fitting and multiple-planet search concludes when no additional TCEs are produced or an upper limit is reached.

The algorithms presented in this chapter build on the work of Tenenbaum et al. (2010) and Wu et al. (2010). The performance of the transit model fitting and multiple-planet search is demonstrated in Section 12.8. Finally, conclusions are presented in Section 12.9.

## 12.2 Light Curve Preprocessing

The light curves of target stars are processed in the PDC module before they are input in the DV module. As described in Chapter 8, the systematic errors due to the thermal transients and optical distortions are estimated and compensated, outliers due to cosmic rays and transients due to argabrightenings are removed, and sudden pixel sensitivity dropouts (SPSDs) are identified and corrected. Nevertheless, the light curves from PDC have to be preprocessed further to be ready for the transit model fitting. The preprocessing procedures in DV include baseline removal, light curve normalization, quarterly data segment stitching, harmonic removal, and timestamp conversion.

### 12.2.1 Baseline Removal and Light Curve Normalization

The light curve generated in the PDC module measures the absolute brightness value of the target star in units of photo-electrons ( $e^-$ ). Since the brightness of one target star is generally measured by four different CCD channels due to the quarterly rotations of the spacecraft about the telescope boresight, the baseline of the measured light curve of the target star varies from quarter to quarter. The preprocessing procedure of baseline removal and light curve normalization removes the baseline of the measurement and generates the normalized light curves so that they can later be uniformly processed by the transit model fitting algorithms. This preprocessing procedure is implemented quarter by quarter in two steps: 1) For each target star, the median flux level is determined for each quarter. For stars on the same CCD channel, the median flux level varies from one target star to another depending on the magnitude and spectrum of the target star; for a given target star, the median varies from quarter to quarter depending on the sensitivity of the CCD pixels and the sub-pixel location of the stellar image. 2) The median is subtracted from the corresponding quarterly light curves and the difference is normalized by the median, yielding a normalized light curve in units of parts per million (ppm). For the out-of-transit data points, the baseline value is zero. For in-transit data points, the normalized flux is negative and its absolute value corresponds to the ratio of the flux blocked by the transiting planet to the total flux of the unblocked target star. For example, to an extraterrestrial observer, the depth of the normalized light curve of the Earth transiting the Sun is about  $84 \times 10^{-6}$ , or 84 parts ppm.

### 12.2.2 Quarterly Data Segment Stitching

The light curve of a target star is separated into segments with gaps that may have resulted from the the quarterly rolls, the monthly data downlinks, or spacecraft anomalies. The preprocessing procedure of data segment stitching removes the trend and transient of the light curve and fills the gaps between the segments. The trend of the light curve of each segment is identified, and the light curve at the edges of the segments, where transients are usually observed, is fitted with a model of linear and exponential components. Then the detrending algorithm removes the identified trend and the fitted components. The gaps between the data segments are filled with different methods, depending on the length of the gaps: the short gaps are filled with an auto-regressive model and the long gaps are filled via data reflection and tapering (see Chapter 9 for more information about the data gap filling algorithms).

### 12.2.3 Harmonic Removal

The harmonic removal procedure identifies and removes sinusoidal harmonic components, which are significant in the light curves of target stars such as contact binaries. The light curve is first processed with a Fast Fourier Transform (FFT) to determine the frequencies of the significant harmonic components. Then the magnitude and phase of the components are fitted and the significant harmonic components are removed from the light curve. The harmonic removal algorithm is discussed in detail in Chapter 9.

### 12.2.4 Timestamp Conversion

Based on Kepler's laws of planetary motion, the transits of exoplanets are inherently periodic<sup>2</sup> if the observer is located at the barycenter of the solar system and the events are measured in the Barycentric Dynamic Time (BDT) frame. However, the timestamps associated with the input PDC light curves are Modified Julian Day (MJD) numbers, which correspond to the time when the light of the events of the target star arrives at the *Kepler* spacecraft in the Coordinated Universal Time (UTC) frame. Before the transit model fitting algorithms are applied, barycentric time corrections are applied to obtain timestamps in Barycentric Modified Julian Day (BMJD) numbers, to correspond to the time when the light of the events of the target star arrives at the barycenter of the solar system in the BDT frame.

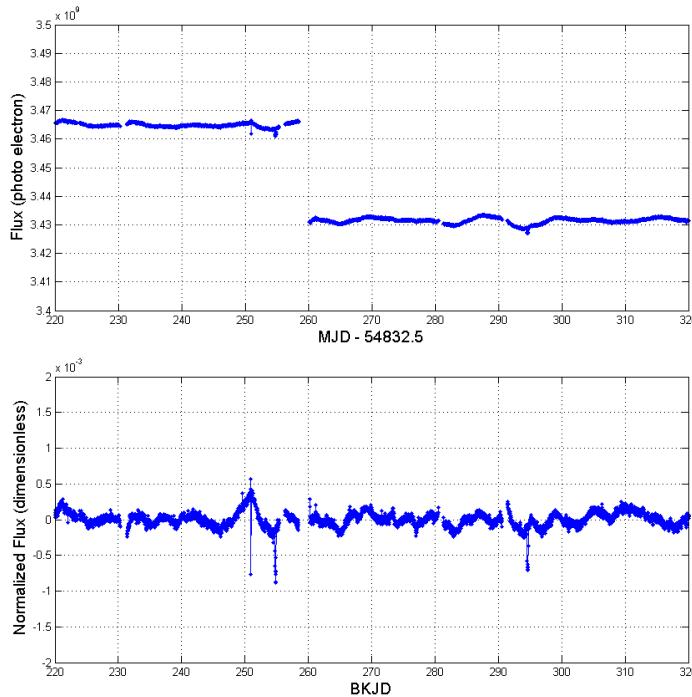
The algorithm to determine the BMJD timestamp first collects the following inputs: the ephemeris of the *Kepler* spacecraft, the ephemeris of the solar system, and the celestial coordinates of the target star. Then the difference between the time when the light of the events of the target star reaches the barycenter of the solar system and the time when the same light arrives at the *Kepler* spacecraft, which is located in an Earth-trailing heliocentric orbit, is calculated. Finally, the BMJD timestamps are determined as the sum of the MJD timestamps and the aforementioned barycentric time corrections. To simplify the processing and storage of the *Kepler* science data, a new timestamp, Barycentric *Keplerian* Julian Date (BKJD), is defined and used in the *Kepler* Science Data Processing Pipeline and the NASA Exoplanet Archive. By definition, BKJD is equal to BMJD minus a constant of 54,832.5, which corresponds to 12:00:00 noon on January 1, 2009 (the first day of the year when the *Kepler* spacecraft was launched). After the preprocessing procedure of timestamp conversion, all light curves are attached to new timestamps of BKJD. The time frames and the timestamps before and after timestamp conversion in the preprocessing are summarized in Table 12.1.

**Table 12.1** Time frames and timestamps before and after the timestamp conversion.

Before/After Conversion	Time Frame	Timestamp
before	UTC	MJD
after	BDT	BKJD (=BMJD-54,832.5)

As an example, Figure 12.3 shows the light curves of the target star KIC 8478994, or *Kepler*-37, before and after the preprocessing procedures. As illustrated in the figure, the light curve before the preprocessing shows the absolute flux value in units of photo-electrons, timestamped in MJD, and the light curve after the preprocessing shows the dimensionless normalized flux value, timestamped in BKJD.

<sup>2</sup>This neglects transit timing variations, which can be quite large for dynamically packed planetary systems with planets in near-orbital resonances.



**Figure 12.3** Flux time series of KIC 8478994 before (top) and after (bottom) the preprocessing procedures.

### 12.3 Geometric Transit Model

The transit model fitting algorithms of the DV module basically use a geometric transit model of Mandel & Agol (2002), including a nonlinear limb-darkening model, parameterized as per Claret & Bloemen (2011). The limb-darkening depends on the stellar parameters, such as the radius  $R_s$  (solar radii), surface gravity  $\log g$  ( $\log_{10} (cm s^{-2})$ ), metallicity  $\log_{10} [M/H]$  (dimensionless), and effective temperature  $T_{eff}$  ( $K$ ), which are extracted from the Kepler Input Catalog (KIC).

#### 12.3.1 Fitted Parameters

In the geometric transit model, the eccentricity and the longitude of periapsis of the planet orbit around the host star are assumed to be 0, and the five parameters to be fitted are listed below:

- Transit epoch time  $t_{epoch}$  (BKJD): the time corresponding to the center of the first detected transit;
- Orbital period  $P$  (days): the interval between consecutive planetary transits, i.e., the period of the planet's orbit;
- Ratio of planet radius to stellar radius  $R_p/R_s$  (dimensionless): the ratio of the planet radius divided by the stellar radius;
- Ratio of semi-major axis to stellar radius  $a/R_s$  (dimensionless): since the eccentricity of the planet's orbit is assumed to be zero, this is the ratio of the distance between the planet and the star divided by the stellar radius;
- Impact parameter  $b$  (dimensionless): the sky-projected distance between the center of the stellar disc and the center of the planet disc at conjunction, normalized by the stellar radius.

As discussed in Subsubsection 12.5.1.2, the fitted parameters are determined with the iterative Levenberg-Marquardt (LM) algorithm. The initial values of the fitted parameters are determined from the TCE parameters provided by the TPS module. The TPS values for orbital period can be used directly. Note that the transit epoch time from the TPS module is in units of MJD, while the fitted parameter of  $t_{epoch}$  is in units of BKJD; therefore, a unit conversion is required. The impact parameter  $b$  is initially set to a fixed value between 0 and 1 (the default value is 0.1). The initial values of  $R_p/R_s$  and  $a/R_s$  are determined from the TCE parameters according to:

$$\frac{R_p}{R_s} = \sqrt{\frac{SES_{TPS}}{SNR} \sqrt{\frac{d_{lc}}{d_{tr,TPS}}}} \text{ and} \quad (12.1)$$

$$\frac{a}{R_s} = \sqrt{\frac{\left(1 + \frac{R_p}{R_s}\right)^2 - b^2}{\sin^2\left(\frac{\pi d_{tr,TPS}}{24 P_{TPS}}\right)} + b^2}, \quad (12.2)$$

where the single event statistic  $SES_{TPS}$  (dimensionless), orbital period  $P_{TPS}$  (days), and transit duration  $d_{tr,TPS}$  (hours) are TCE parameters determined in the TPS module.  $SNR$  is the signal-to-noise ratio (SNR) of the light curve and  $d_{lc}$  (hours) is the duration of a LC interval (29.4 min or 0.49 hr).

### 12.3.2 Derived Parameters

When the transit model fitting algorithm has converged, several additional parameters regarding the planet or the transits can be derived from the fitted parameters.

Given the stellar radius  $R_s$ , the planet radius  $R_p$  is determined directly from the fitted parameter  $R_p/R_s$ :

$$R_p = \left(\frac{R_\odot}{R_\oplus}\right) \left(\frac{R_p}{R_s}\right) R_s, \quad (12.3)$$

where  $R_\odot$  and  $R_\oplus$  are radii of the Sun and the Earth, respectively, both in units of  $m$ . Since  $R_s$  is in units of solar radii,  $R_p$  given by Equation 12.3 is in units of Earth radii.

Before calculating the semi-major axis of the planet orbit  $a$ , the planet-star separation for a circular orbit, the acceleration due to gravity at the surface of the star  $g$  should first be determined from the stellar parameter  $\log g$  as:

$$g = \frac{1}{100} 10^{\log g}. \quad (12.4)$$

A factor of 1/100 is required to get the acceleration  $g$  in units of  $m s^{-2}$  from  $\log g$  in units of  $\log_{10}(cm s^{-2})$ .

The semi-major axis of the planetary orbit  $a$  is not determined directly from the fitted parameter  $a/R_s$ , but derived from the fitted orbital period  $P$  based on Kepler's third law:

$$a = \frac{1}{f_{AU}} \left[ \frac{(86400 P) (R_s R_\odot) \sqrt{g}}{2\pi} \right]^{\frac{2}{3}}, \quad (12.5)$$

where  $f_{AU}$  is the factor to convert the astronomical unit (AU) to  $m$ . Please note  $P$  and  $R_s$  are in units of days and solar radii, respectively, so Equation 12.5 gives the semi-major axis of the planet orbit,  $a$ , in AU.

The inclination of the planet orbit  $i$  in units of degrees, is determined from fitted parameters  $b$  and  $a/R_s$ :

$$i = \frac{180}{\pi} \cos^{-1} \left( \frac{b}{a/R_s} \right). \quad (12.6)$$

As illustrated in Figure 12.4, the transit depth  $D$ , transit duration  $d_{tr}$ , and transit ingress time  $d_{in}$  are another set of parameters defining the size and shape of a transit. The transit depth  $D$  is determined as the absolute value of the minimum of the normalized light curve generated from the fitted parameters (to be discussed in Section 12.4), multiplied by a factor of  $10^6$  to convert the dimensionless normalized flux value to the transit depth in units of ppm. The parameters  $d_{tr}$  and  $d_{in}$ , both in units of hours, are derived from fitted parameters  $R_p/R_s$ ,  $a/R_s$ ,  $b$ , and  $P$  with the following equations:

$$d_{tr} = \frac{24 P}{\pi} \sin^{-1} \left( \sqrt{\frac{\left(1 + \frac{R_p}{R_s}\right)^2 - b^2}{\left(\frac{a}{R_s}\right)^2 - b^2}} \right) \text{ and} \quad (12.7)$$

$$d_{in} = \frac{12 P}{\pi} \sin^{-1} \left( \sqrt{\frac{\left(1 + \frac{R_p}{R_s}\right)^2 - b^2}{\left(\frac{a}{R_s}\right)^2 - b^2}} - \sqrt{\frac{\left(1 - \frac{R_p}{R_s}\right)^2 - b^2}{\left(\frac{a}{R_s}\right)^2 - b^2}} \right). \quad (12.8)$$

The planet equilibrium temperature  $T_{eq}$ , an approximation for the temperature of the planet, is calculated assuming a thermodynamic equilibrium is reached between the incident stellar flux and the radiated heat from the planet:

$$T_{eq} = T_{eff} (1 - \alpha)^{\frac{1}{4}} \sqrt{\frac{R_s R_\odot}{2 a f_{AU}}}, \quad (12.9)$$

where  $a$  is the semi-major axis of the planetary orbit in AU determined by Equation 12.5,  $\alpha$  is the albedo of the planet, whose default value is set to 0.3, and both  $T_{eff}$  and  $T_{eq}$  are in  $K$ .

The planet effective stellar flux,  $\phi_{eff}$ , defined as the ratio of the flux of the host star at the planet to the solar flux at the Earth, is calculated as

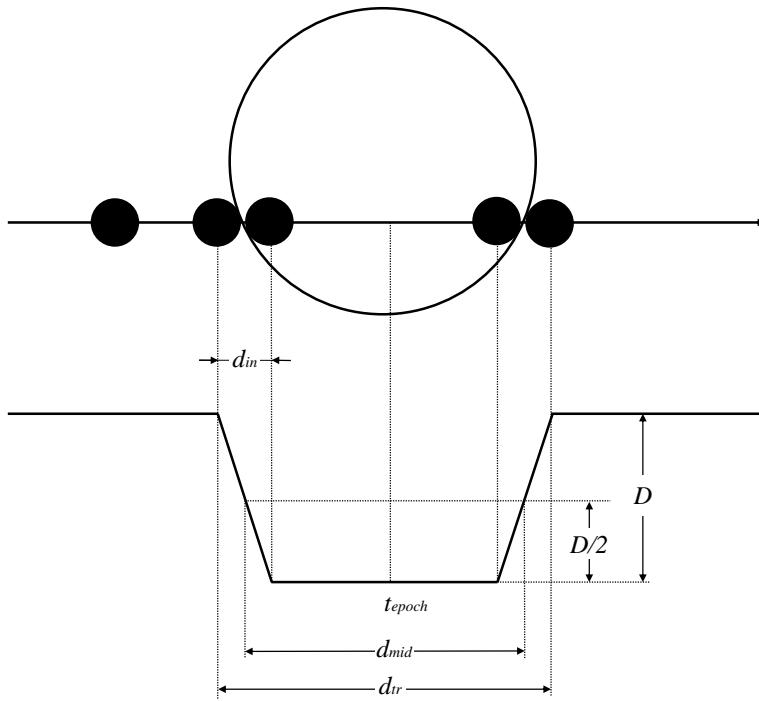
$$\phi_{eff} = \left( \frac{R_s}{a} \right)^2 \left( \frac{T_{eff}}{T_{eff,\odot}} \right)^4, \quad (12.10)$$

where  $a$  is determined by Equation 12.5 and  $T_{eff,\odot}$  is the effective temperature of the Sun in units of  $K$ .

The fitted and derived parameters of the transit model are summarized in Table 12.2.

## 12.4 Geometric Transit Signal Generator

The geometric transit signal generator generates a light curve at an input array of timestamps of cadences in BKJD (nominally the timestamps corresponding to the midpoints of cadences)



**Figure 12.4** Schematic of planetary transit and associated light curve with the depth, duration, ingress time, and epoch time indicated.

**Table 12.2** Fitted and derived parameters of the transit model.

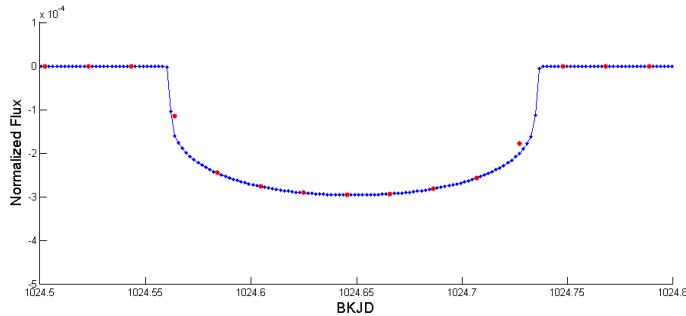
Parameter	Symbol	Unit	Fitted/Derived
transit epoch time	$t_{epoch}$	BKJD	fitted
planet orbital period	$P$	day	fitted
ratio of planet radius to star radius	$R_p/R_s$	dimensionless	fitted
ratio of semi-major axis to star radius	$a/R_s$	dimensionless	fitted
impact parameter	$b$	dimensionless	fitted
planet radius	$R_p$	Earth radius	derived
planet orbit semi-major axis	$a$	AU	derived
planet orbit inclination	$i$	degree	derived
transit depth	$D$	ppm	derived
transit duration	$d_{tr}$	hour	derived
transit ingress time	$d_{in}$	hour	derived
planet equilibrium temperature	$T_{eq}$	$K$	derived
planet effective stellar flux	$\phi_{eff}$	dimensionless	derived

with the fitted parameters of a geometric transit model described in Subsection 12.3.1. The coefficients of the limb-darkening model are determined by the stellar parameters of the target star.

The computation of the light curve is implemented in the following steps. First, an array of oversampled timestamps is constructed from the input array of timestamps. This is necessary in order to obtain an accurate flux level estimate at the temporal resolution of the data (29.4 min).

For each element in the input array of timestamps, a sub-array of 11 oversampled timestamps is generated. The step size of the oversampled timestamps is 1/11 of an LC interval or 2.67 min. The center element of the sub-array, the 6th of the 11 elements, is equal to the corresponding element in the input array of timestamps. The oversampled timestamp array is determined as the union of all sub-arrays generated with all input timestamps. Then the oversampled timestamps that fall within a transit (including a small buffer on each side of the transit) are identified with the parameters  $t_{epoch}$  and  $P$ . A circular Keplerian orbit, normalized by the stellar radius  $R_s$ , is determined from the parameters  $a/R_s$  and  $b$ . The positions of the planet in the orbit, vectors pointing from the target star to the planet, are computed and the corresponding impact parameters are determined by projecting the position vectors to the plane perpendicular to the direction of the target star. The relative flux value, the ratio of the stellar flux blocked by the transiting planet to the unblocked stellar flux, is calculated for each oversampled timestamp with the impact parameter  $b$ , the fitted parameter  $R_p/R_s$ , and the limb-darkening coefficients. Finally, the relative flux at each of the input timestamps is determined as the mean of 11 relative flux values at the corresponding oversampled timestamps.

Figure 12.5 shows the normalized flux time series generated by the geometric transit signal generator with following parameters:  $t_{epoch} = 138.50000$  days,  $P = 10.30405$  days,  $R_p/R_s = 0.0155697$ ,  $a/R_s = 18.7471$ , and  $b = 0.1$ , which are determined by the reduced parameter fit (to be discussed in Subsection 12.5.2) of the 6th TCE of the target star KIC 6541920, also known as the planet *Kepler-11b*. As shown in the figure, the step size of the input timestamps is the duration of an LC (29.4 min), and the normalized flux values at the input and oversampled timestamps are plotted as red and blue dots, respectively.



**Figure 12.5** Normalized light curves of the 6th TCE of KIC 6541920 generated by the geometric transit signal generator. The normalized flux values at the oversampled timestamps, whose step size is 1/11 of an LC interval, or 2.67 min, are plotted as blue dots. Each of the normalized flux values at the LC interval, or 29.4 min, is determined as the mean of 11 corresponding values at oversampled timestamps and plotted as a red dot.

Since the surface brightness of a star appears to vary due to the limb-darkening effect, the calculation of the normalized flux value is implemented with an iterative numerical integration algorithm. At each iteration, the integration step is cut into half and an updated normalized flux time series is determined with the nonlinear limb-darkening model. The iterative process is terminated when the required precision is satisfied or when an upper limit of the execution time of the iterative algorithm is reached. If the parameter  $R_p/R_s$  is less than 0.01, a small-body approximation is used to speed up the algorithm, assuming the stellar surface brightness is constant under the disc of the eclipsing object.

The five fitted parameters defined in Subsection 12.3.1 can be divided in two relatively independent groups: 1) the transit epoch time  $t_{epoch}$  and the orbital period  $P$  define the occurrence time of the transits; 2) the ratio of planet radius to star radius  $R_p/R_s$ , the ratio of semi-major

axis to star radius  $a/R_s$ , and the impact parameter  $b$ , define the depth, duration, and shape of the transits.

Figure 12.6 illustrates how the variations of the parameters  $R_p/R_s$ ,  $a/R_s$ , and  $b$  change the depth, duration, and shape of the transits in the light curves. As the reference for comparison, the light curve shown in Figure 12.5 is plotted as blue in Figure 12.6. The corresponding parameter values are used as references for the parameter variations. When  $R_p/R_s$  is increased by 10% and 20% to its reference value, the corresponding model light curves are plotted as red and black lines, respectively, in the plot on the top of Figure 12.6. Since  $R_p/R_s$  defines the relative size of the transiting planet to the host star, an increase of  $R_p/R_s$ , meaning more stellar flux is blocked by the transiting planet, leading to an increase of the transit depth. When  $a/R_s$  is increased by 10% and 20% to its reference value, the corresponding model light curves are plotted as red and black lines, respectively, in the middle plot of Figure 12.6. Since the orbital period,  $P$ , is fixed, an increase of  $a/R_s$ , indicating a decrease of the stellar radius,  $R_s$ , leads to a decrease of the transit duration. When  $b$  is changed to 0.3, 0.5, 0.7, and 0.9, the corresponding model light curves are plotted as red, black, magenta, and green lines, respectively, in the plot on the bottom of Figure 12.6. An increase of  $b$  moves the transit trajectory toward the edge of the stellar disc and results in a decrease of the transit duration. Since the point at the center of the transit moves toward the edge of the stellar disc, the transit depth decreases as well due to the limb-darkening effect.

## 12.5 Geometric Model Fitting Algorithms

The inputs of the geometric model fitting algorithms include two data sets: 1) the light curve after the preprocessing procedures described in Section 12.2 and 2) the TCE parameters, including transit epoch time, orbital period, trial transit duration, and single and multiple event statistics, generated by the TPS module.

Subsection 12.5.1 discusses an iterative whitening and model fitting process, used in the all-transit fit, odd-even transit fit, and reduced-parameter fit. The TCE parameters are used to seed the initial values of reduced parameter fits. Subsection 12.5.2 describes a fitting algorithm to resolve the degenerate problem of fitting the impact parameters. Subsection 12.5.3 describes the algorithms to fit odd and even transits, whose outputs are used in the diagnostic test to distinguish the transiting planets from eclipsing binaries.

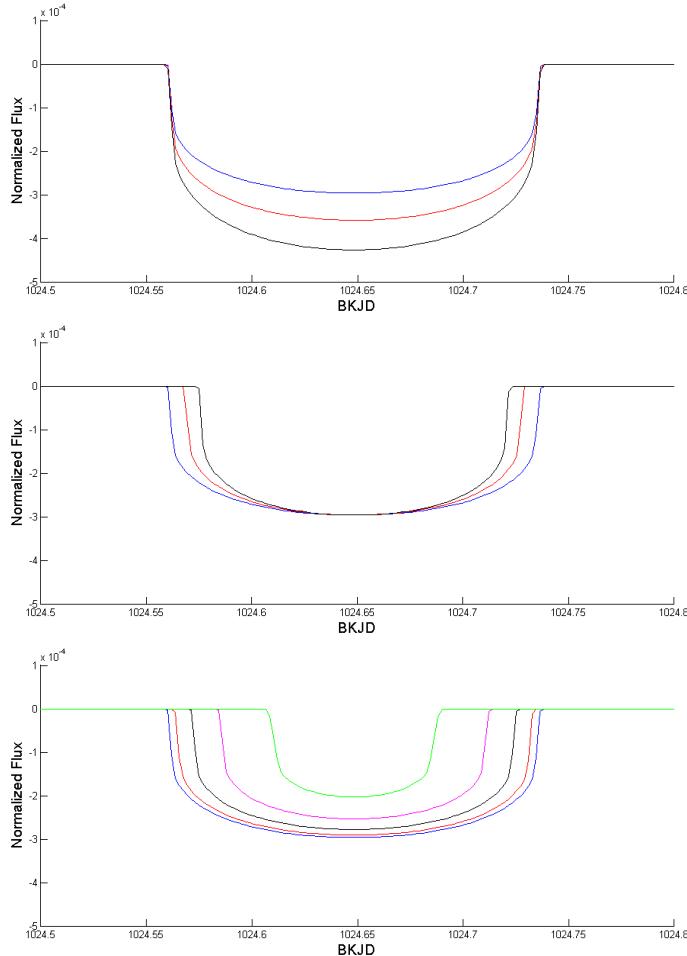
When the fitting algorithm completes successfully, the outputs of the fitter are described in Subsection 12.5.4. When the fitting algorithm fails, the alert messages are generated and discussed in Subsection 12.5.5.

### 12.5.1 Iterative Whitening and Model Fitting

Compared to the transit features, secular variations due to pointing drift, focus variations, and stellar variability can be quite large. The secular variations of the light curve, appearing as correlated noise, can lead to biases in the fitted parameters of the geometric transit model. Therefore, a whitening filter is applied to the light curves before the transit model fitting to account explicitly for the correlation structure of the noise. Considering that the whitening filter changes the shape of the transits, the same whitening filter is applied to the model light curve generated by the geometric transit signal generator.

The flux time series of a target star at times  $t_i$ ,  $i = 1, 2, \dots, N$ , is denoted as  $y(t_i)$ . Let  $\theta$  denote a  $5 \times 1$  vector of fitted parameters:

$$\theta = [t_{epoch} \quad P \quad R_p/R_s \quad a/R_s \quad b]^T. \quad (12.11)$$



**Figure 12.6** Light curves generated by the geometric transit signal generator with different parameters of  $R_p/R_s$  (top),  $a/R_s$  (middle), and  $b$  (bottom). See text for an explanation.

The predicted light curve generated from the geometric transit model with the parameter vector  $\theta$  is denoted as  $s(t_i, \theta)$ . When the whitening filter is applied to the time series  $y(t_i)$  and  $s(t_i, \theta)$ , the corresponding whitened time series are denoted as  $\tilde{y}(t_i)$  and  $\tilde{s}(t_i, \theta)$ , respectively. The geometric transit model fitting is implemented with a LM algorithm to search for the vector  $\theta$  in the parameter space to minimize the following weighted nonlinear least-squares cost function:

$$\chi^2(\theta) = \sum_{i=1}^N w_i [\tilde{y}(t_i) - \tilde{s}(t_i, \theta)]^2 \quad (12.12)$$

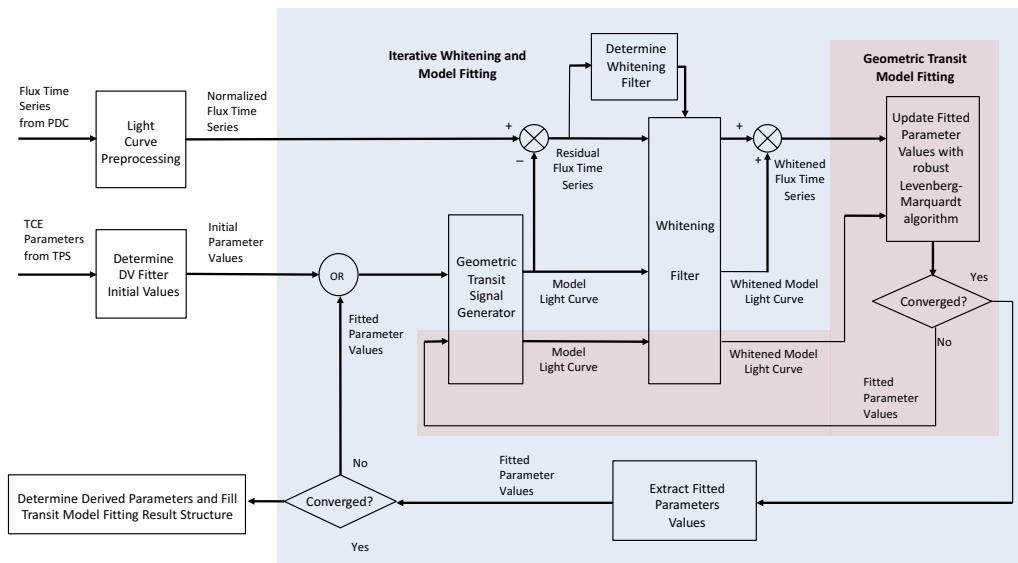
where  $w_i, i = 1, 2, \dots, N$  are weights, ranging between 0 and 1. During the fit, these weights are adjusted to deemphasize points with large departures from the model values, in order to reduce the impact of outliers.

Let  $\tilde{\mathbf{y}}$  and  $\tilde{\mathbf{s}}(\theta)$  denote vectors of measured and predicted light curves in the whitened domain, respectively, and  $\mathbf{W}$  denote a diagonal matrix of the weights. Equation 12.12 can be rewritten in the following matrix form:

$$\chi^2(\theta) = [\tilde{\mathbf{y}} - \tilde{\mathbf{s}}(\theta)]^T \mathbf{W} [\tilde{\mathbf{y}} - \tilde{\mathbf{s}}(\theta)]. \quad (12.13)$$

Since the out-of-transit light curve data just show the measurement noise around the baseline value of zero, they offer no information to characterize the transits. Therefore, the transit model fitting is restricted to the data within or close to the transits. The center times of the transits are predicted from the parameters  $t_{epoch}$  and  $P$ , and only the light curve data whose timestamps fall in the time ranges of 5 times the transit duration, centered at the transit center, are used in the geometric transit model fit. The data selection can also be viewed as a fitter implemented to all data points with different weights: the weight is set to 1 in Equation 12.12 when the difference between the timestamp of the data point and the center time of the nearest transit is less than 2.5 times the transit duration; otherwise, the weight is set to 0.

For each TCE generated by TPS, the geometric transit model fitting is implemented with a loop that includes an adaptive whitening filter and a robust LM transit fitter, as shown in Figure 12.7. The whitening filter turns the correlated noise in the measured flux time series to an uncorrelated, or white, noise. The predicted light curve is subjected to the same whitening filter, so the fitted parameters of the geometric transit model are determined by a nonlinear least-squares fitting in the whitened domain. The fit residual is utilized to update the parameters of the whitening filter on each iteration. Robust weights are assigned to each point of the flux time series so that data with large errors are assigned small weights in the nonlinear least-squares fitting algorithm. The iterative whitening and fitting loop is terminated when both the whitening filter and the transit fitter converge or a predefined iteration limit is reached.



**Figure 12.7** Block diagram of the iterative whitening and model fitting process. Two loops are shown in the figure: the outer loop, shaded in blue, includes a whitening filter and a transit model fitter, and the parameters of the whitening filter are updated on each iteration with the residuals of the transit model fitter; the inner loop, shaded in pink, includes the LM fit and robust weight reassignment.

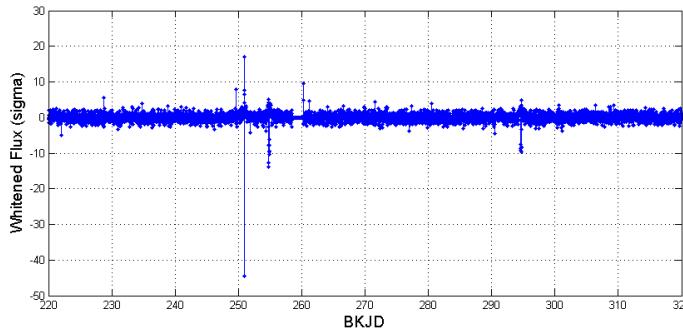
**12.5.1.1 Whitening filter** Considering the non-stationary nature of the stellar variability, an adaptive whitening filter is generated and used to remove variations in the light curve.

The whitening filter is implemented in the following steps: 1) The flux time series and the model transit light curves are mapped into a two dimensional array of whitening coefficients,

localizing the signal both in time and in frequency with the Overcomplete Wavelet Transform (OWT), a modified version of the discrete wavelet transform (Jenkins, 2002); 2) The noise power in each wavelet bandpass is estimated using a moving median absolute deviation (MAD) filter; 3) The wavelet coefficients of the flux time series and the model transit light curve are normalized by the root-mean-square (rms) noise power estimates. Finally, the whitened time series are reconstructed from the updated wavelets with the inverse OWT. The whitening filter is discussed in detail in Chapter 9.

A consequence of the whitening filter is that the shape of the transit in the measured flux time series is distorted by the filter; it is therefore necessary to apply the same filter to the predicted model light curve. The fitted parameters of the geometric transit model are determined by the nonlinear least-squares fitting in the whitened domain.

Figure 12.8 shows the whitened flux time series of KIC 8478994 in an interval of 100 days, which is the output generated when the unwhitened normalized flux time series, shown on the bottom of Figure 12.3 in Section 12.2, goes through a whitening filter. Figure 12.9 illustrates the same unwhitened and whitened flux time series in an interval of 6 days, so that the distortion of the whitening filter on the shape of the transit is shown. It can be seen that the depth of the transit is about  $6 \times 10^{-4}$ , or 600 ppm, while the stellar variability causes the variations of more than  $3 \times 10^{-4}$  in the unwhitened flux time series. The whitened flux time series, whose standard deviation is equal to 1, is in units of standard deviations of the unwhitened flux.

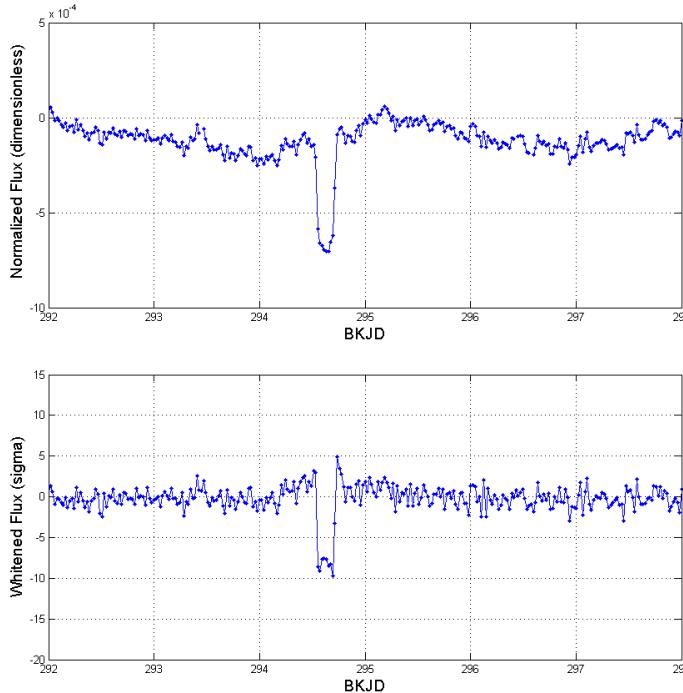


**Figure 12.8** Whitened flux time series of KIC 8478994.

**12.5.1.2 LM fit of geometric transit model parameters** The LM algorithm is implemented to search for parameters that minimize the nonlinear least-squares cost function, such as the  $\chi^2$  metric defined in Equation 12.12 or Equation 12.13.

In the general form of the LM algorithm, there is no restriction for the values that the fitted parameters can take. However, in the geometric transit model, the parameters of  $P$ ,  $R_p/R_s$ ,  $a/R_s$ , and  $b$  only take positive values. Therefore, in the geometric transit model fitting algorithms, all of the fitted parameters are forced to be positive values. When an updated value of a parameter turns out to be negative in the search process, the parameter is reset to the absolute value of the updated value so that all the fitted parameters are positive.

An additional subtlety to the parameterization is that the impact parameter is constrained to lie in the range  $[0, 1]$  but the LM algorithm implicitly requires all fit parameters to be valid over all real values. To address this mismatch, a nonlinear transformation in the form of a sin function is performed between the “internal” parameter used by the LM algorithm and the “external parameter” used in the geometric transit model; this transformation maps the range  $[-\infty, \infty]$  used by the LM algorithm to the range  $[-1, 1]$  for the impact parameter in the geometric transit model, and the negative value is updated with its absolute values, as discussed above.

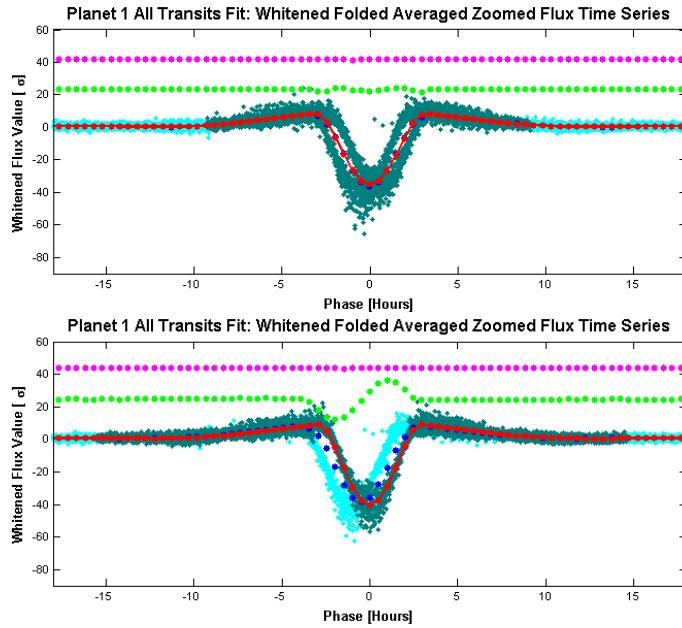


**Figure 12.9** Flux time series of KIC 8478994 before (top) and after (bottom) a whitening filter is applied.

In the SOC 9.3 codebase for DR25, the iterative search process of the parameter vector  $\theta$  will stop if the relative variation of the  $\chi^2$  metric is less than 0.1%, or the absolute values of the differences of the fitted parameters are less than 10% of the corresponding uncertainties, or a preset limit of 100 iterations is reached. The threshold values are set as module parameters in the DV input structure.

**12.5.1.3 Robust fit** In the weighted nonlinear least-squares fitting problem given by Equation 12.12 or Equation 12.13, the weight of each data point used in the fit is initially set to either 1 or 0, depending on the data selection, or whether the timestamp of the data point is within 2.5 times the transit duration from the center time of the nearest transit. However, when some of the selected data points are outliers, the fitting algorithm tries to reach a compromised solution between the good data points and outliers, usually resulting in biases in the fitted parameters.

The robust fitting algorithm, which is optional, works by assigning a weight in the range between 0 and 1 to each data point for the fit. The outliers are assigned to the weights close to 0 so that the output of the robust fitting algorithm is less sensitive to the outliers in the data. The robust fitting algorithm is executed after the convergence of the LM fit with all weights set to 1 or 0; the weights are re-assigned and the LM fit is done iteratively. In each iteration, the weight of each data point is calculated from the fit residual of the previous iteration with a bisquare function, so that the data points with larger residuals are assigned smaller weights. The iterative process of weight re-assignment and LM fit continues until the fitted parameters converge within a specified tolerance. The effect of robust fit is demonstrated with the following example. The primary and secondary eclipses of an eclipsing binary target (KIC 6960446) are identified as one TCE by the TPS module. Figure 12.10 shows the folded flux time series of the target and the folded model light curve generated with the fitted parameters of the TCE, both in the whitened domain, when the robust fit is off (top) and on (bottom), respectively. The secondary eclipse, which has a smaller depth, is located at phase 0 of the plot. A small phase offset is observed



**Figure 12.10** Folded flux time series of KIC 6960446 and model light curve generated with fitted parameters of the TCE, both in the whitened domain, when a) robust fit is disabled (top) and b) robust fit is enabled.

in the folded primary and secondary eclipses. In the plot, the flux data points are illustrated as dark green dots when the weights of the fit are larger than 0.1, otherwise, illustrated as light blue dots. When the weights of the fit are equally set to 1 for the data points, the fitted model light curve compromises between the primary and secondary eclipses. However, the weights of the fit are calculated iteratively in the robust fitting algorithm. As a result, most data points of the primary eclipses are identified as outliers and assigned weights less than 0.1 at the end of the iterative process, and the robust fit algorithm generates unbiased fitted parameters to characterize the secondary eclipses.

**12.5.1.4 Goodness of Fit Metrics** Two sets of goodness of fit metrics are calculated when the transit model fitting algorithm is completed successfully. One set includes the  $\chi^2$  metric and the number of degrees of freedom, the other set is the SNR of the fit.

The  $\chi^2$  metric is determined with Equation 12.12, and the number of degrees of freedom is determined as the sum of the weights minus the number of fitted parameters. It is noted the weights take values of either 0 or 1 when the robust fit is disabled, as described in Subsubsection 12.5.1.3.

The SNR of the fit is determined as:

$$SNR_{fit} = \sqrt{\tilde{s}(\hat{\theta})^T W \tilde{s}(\hat{\theta})}, \quad (12.14)$$

where  $\hat{\theta}$  is the vector of fitted parameters and  $\tilde{s}(\hat{\theta})$  is the whitened model light curve generated with  $\hat{\theta}$ .

The  $\chi^2$  metric and the number of degrees of freedom shows the distance between the flux time series and the model light curve in the whitened domain. The SNR shows the strength of the TCE relative to the noise.

**12.5.1.5 Uncertainties of fitted and derived parameters** Let  $\mathbf{H}$  denote the Jacobian of the model light curve  $\tilde{\mathbf{s}}(\theta)$  to the vector of fitted parameters  $\theta$ , such that:

$$\mathbf{H} = \frac{\partial \tilde{\mathbf{s}}(\theta)}{\partial \theta}. \quad (12.15)$$

The covariance matrix of the fitted parameters is determined as

$$\text{Cov}(\hat{\theta}) = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \sigma_{res}, \quad (12.16)$$

where  $\sigma_{res}$  is the root of the mean squared (rms) value of the residuals of the fit. The elements of the Jacobian  $\mathbf{H}$  are determined numerically.

Let  $\alpha$  and  $\psi$  denote vectors of stellar parameters and derived parameters, as defined below:

$$\alpha = [R_s \ g \ T_{eff}]^T \text{ and} \quad (12.17)$$

$$\psi = [R_p \ a \ i \ d_{tr} \ d_{in} \ D \ T_{eq} \ \phi_{eff}]^T, \quad (12.18)$$

where  $g$  is the acceleration due to gravity at the surface of the star, determined from the stellar parameter  $\log g$  as shown in Equation 12.4. The uncertainty of  $g$  ( $m s^{-2}$ ) can be determined from the uncertainty of  $\log g$  ( $\log_{10} (cm s^{-2})$ ) multiplied by  $g \ln 10$ .

As discussed in Subsection 12.3.2,  $\psi$  is a function of  $\theta$  and  $\alpha$ . The covariance matrix of  $\psi$ ,  $\text{Cov}(\psi)$ , includes the components propagated both from the covariance matrix of  $\theta$ ,  $\text{Cov}(\theta)$ , and the uncertainties of the elements of  $\alpha$ , as shown below:

$$\text{Cov}(\psi) = \left( \frac{\partial \psi}{\partial \theta} \right)^T \text{Cov}(\theta) \left( \frac{\partial \psi}{\partial \theta} \right) + \left( \frac{\partial \psi}{\partial \alpha} \right)^T \text{Cov}(\alpha) \left( \frac{\partial \psi}{\partial \alpha} \right) \quad (12.19)$$

where  $\text{Cov}(\alpha)$  is a diagonal matrix whose elements are squares of the uncertainties of the corresponding stellar parameters.

The Jacobians  $\partial \psi / \partial \theta$  and  $\partial \psi / \partial \alpha$  have the following forms:

$$\frac{\partial \psi}{\partial \theta} = \begin{bmatrix} 0 & 0 & \frac{\partial R_p}{\partial (R_p/R_s)} & 0 & 0 \\ 0 & \frac{\partial a}{\partial P} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial i}{\partial (a/R_s)} & \frac{\partial i}{\partial b} \\ 0 & \frac{\partial d_{tr}}{\partial P} & \frac{\partial d_{tr}}{\partial (R_p/R_s)} & \frac{\partial d_{tr}}{\partial (a/R_s)} & \frac{\partial d_{tr}}{\partial b} \\ 0 & \frac{\partial d_{in}}{\partial P} & \frac{\partial d_{in}}{\partial (R_p/R_s)} & \frac{\partial d_{in}}{\partial (a/R_s)} & \frac{\partial d_{in}}{\partial b} \\ 0 & 0 & \frac{\partial D}{\partial (R_p/R_s)} & \frac{\partial D}{\partial (a/R_s)} & \frac{\partial D}{\partial b} \\ 0 & \frac{\partial T_{eq}}{\partial P} & 0 & 0 & 0 \\ 0 & \frac{\partial \phi_{eff}}{\partial P} & 0 & 0 & 0 \end{bmatrix} \text{ and} \quad (12.20)$$

$$\frac{\partial \psi}{\partial \alpha} = \begin{bmatrix} \frac{\partial R_p}{\partial R_s} & 0 & 0 \\ \frac{\partial a}{\partial R_s} & \frac{\partial a}{\partial g} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{\partial T_{eq}}{\partial R_s} & \frac{\partial T_{eq}}{\partial g} & \frac{\partial T_{eq}}{\partial T_{eff}} \\ \frac{\partial \phi_{eff}}{\partial R_s} & \frac{\partial \phi_{eff}}{\partial g} & \frac{\partial \phi_{eff}}{\partial T_{eff}} \end{bmatrix}. \quad (12.21)$$

Please note the derived parameters  $i$ ,  $d_{tr}$ ,  $d_{in}$ , and  $D$  are determined independently of the stellar parameters; therefore, their partial derivatives with respect to the stellar parameters are all identically zero.

Since the transit depth  $D$  is determined from the model light curve generated by the geometric transit signal generator, the elements  $\partial D / \partial (R_p/R_s)$ ,  $\partial D / \partial (a/R_s)$ , and  $\partial D / \partial b$  of the Jacobian  $\partial \psi / \partial \theta$  are determined numerically. The other non-zero elements of the Jacobians  $\partial \psi / \partial \theta$ , and  $\partial \psi / \partial \alpha$  are calculated according to the following equations:

$$\frac{\partial R_p}{\partial (R_p/R_s)} = \frac{R_p}{(R_p/R_s)}, \quad (12.22)$$

$$\frac{\partial a}{\partial P} = \frac{2}{3} \frac{a}{P}, \quad (12.23)$$

$$\frac{\partial i}{\partial (a/R_s)} = \frac{180}{\pi} \frac{b}{(a/R_s)} \frac{1}{\sqrt{(a/R_s)^2 - b^2}}, \quad (12.24)$$

$$\frac{\partial i}{\partial b} = -\frac{180}{\pi} \frac{1}{\sqrt{(a/R_s)^2 - b^2}}, \quad (12.25)$$

$$\frac{\partial d_{tr}}{\partial P} = \frac{d_{tr}}{P}, \quad (12.26)$$

$$\frac{\partial d_{tr}}{\partial (R_p/R_s)} = \frac{24 P}{\pi} \frac{1 + (R_p/R_s)}{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2} \sqrt{[1 + (R_p/R_s)]^2 - b^2}}, \quad (12.27)$$

$$\frac{\partial d_{tr}}{\partial (a/R_s)} = -\frac{24 P}{\pi} \frac{(a/R_s)}{(a/R_s)^2 - b^2} \frac{\sqrt{[1 + (R_p/R_s)]^2 - b^2}}{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2}}, \quad (12.28)$$

$$\frac{\partial d_{tr}}{\partial b} = -\frac{24 P}{\pi} \frac{b}{(a/R_s)^2 - b^2} \frac{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2}}{\sqrt{[1 + (R_p/R_s)]^2 - b^2}}, \quad (12.29)$$

$$\frac{\partial d_{in}}{\partial P} = \frac{d_{in}}{P}, \quad (12.30)$$

$$\begin{aligned} \frac{\partial d_{in}}{\partial (R_p/R_s)} = \frac{12 P}{\pi} & \left( \frac{1 + (R_p/R_s)}{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2} \sqrt{[1 + (R_p/R_s)]^2 - b^2}} + \right. \\ & \left. \frac{1 - (R_p/R_s)}{\sqrt{(a/R_s)^2 - [1 - (R_p/R_s)]^2} \sqrt{[1 - (R_p/R_s)]^2 - b^2}} \right), \end{aligned} \quad (12.31)$$

$$\begin{aligned} \frac{\partial d_{in}}{\partial (a/R_s)} = -\frac{12 P}{\pi} & \frac{(a/R_s)}{(a/R_s)^2 - b^2} \left( \frac{\sqrt{[1 + (R_p/R_s)]^2 - b^2}}{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2}} - \right. \\ & \left. \frac{\sqrt{[1 - (R_p/R_s)]^2 - b^2}}{\sqrt{(a/R_s)^2 - [1 - (R_p/R_s)]^2}} \right), \end{aligned} \quad (12.32)$$

$$\begin{aligned} \frac{\partial d_{in}}{\partial b} = -\frac{12 P}{\pi} & \frac{b}{(a/R_s)^2 - b^2} \left( \frac{\sqrt{(a/R_s)^2 - [1 + (R_p/R_s)]^2}}{\sqrt{[1 + (R_p/R_s)]^2 - b^2}} - \right. \\ & \left. \frac{\sqrt{(a/R_s)^2 - [1 - (R_p/R_s)]^2}}{\sqrt{[1 - (R_p/R_s)]^2 - b^2}} \right), \end{aligned} \quad (12.33)$$

$$\frac{\partial T_{eq}}{\partial P} = -\frac{1}{3} \frac{T_{eq}}{P}, \quad (12.34)$$

$$\frac{\partial \phi_{eff}}{\partial P} = -\frac{4}{3} \frac{\phi_{eff}}{P}, \quad (12.35)$$

$$\frac{\partial R_p}{R_s} = \frac{R_p}{R_s}, \quad (12.36)$$

$$\frac{\partial a}{R_s} = \frac{2}{3} \frac{a}{R_s}, \quad (12.37)$$

$$\frac{\partial a}{g} = \frac{1}{3} \frac{a}{g}, \quad (12.38)$$

$$\frac{\partial T_{eq}}{R_s} = \frac{1}{6} \frac{T_{eq}}{R_s}, \quad (12.39)$$

$$\frac{\partial T_{eq}}{g} = -\frac{1}{6} \frac{T_{eq}}{g}, \quad (12.40)$$

$$\frac{\partial T_{eq}}{T_{eff}} = \frac{T_{eq}}{T_{eff}}, \quad (12.41)$$

$$\frac{\partial \phi_{eff}}{R_s} = \frac{2}{3} \frac{\phi_{eff}}{R_s}, \quad (12.42)$$

$$\frac{\partial \phi_{eff}}{g} = -\frac{2}{3} \frac{\phi_{eff}}{g}, \text{ and} \quad (12.43)$$

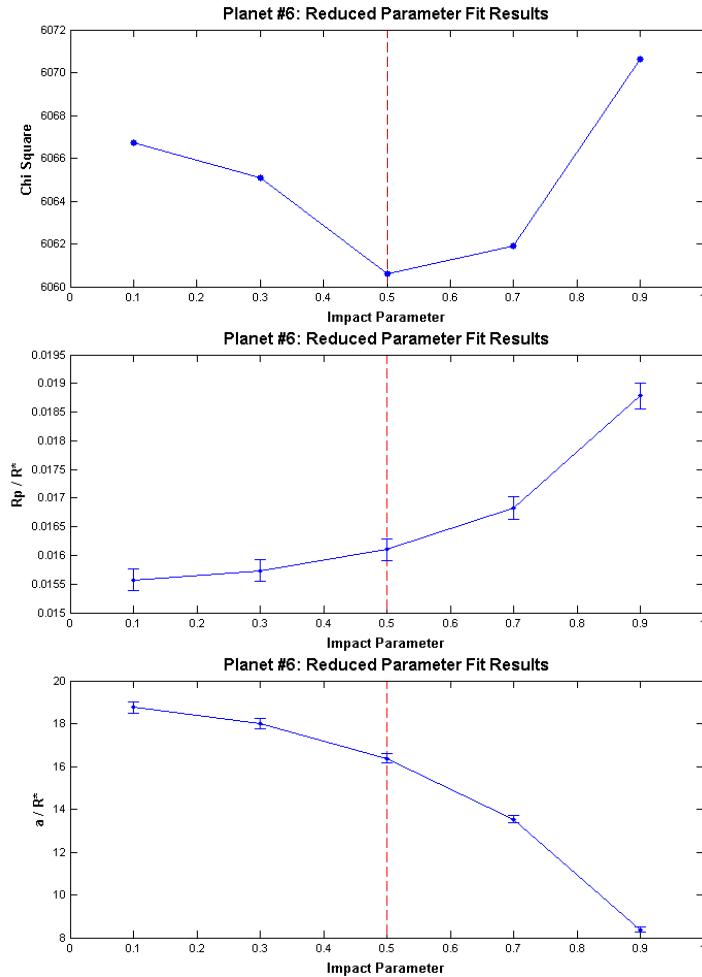
$$\frac{\partial \phi_{eff}}{T_{eff}} = 4 \frac{\phi_{eff}}{T_{eff}}. \quad (12.44)$$

The uncertainties of the fitted and derived parameters, the elements of vectors  $\theta$  and  $\psi$ , are determined as the square roots of the diagonal elements of the matrices  $\text{Cov}(\theta)$  and  $\text{Cov}(\psi)$ , respectively.

### 12.5.2 Reduced Parameter Fits

Out of the five fitted parameters of the geometric transit model defined in Section 12.3, the impact parameter  $b$ , ranging between 0 and 1, basically describes the slope of the edges of transits. When  $b$  is closer to 0, the edges are steeper. Due to the limb-darkening effect of the host star, it is difficult to determine exactly where the transit edges start and end. Therefore, in case of a low SNR for the flux time series, there is insufficient information to determine the impact parameter, which leads to large uncertainties in the fitted parameters. Due to accumulated numerical errors, the outputs of the geometric transit model fit may be inconsistent in different runs of the same code with same input data. To resolve this problem, a set of reduced parameter fits are implemented before the geometric model fitting of all transits and odd-even transits: the impact parameter  $b$  is set to fixed values 0.1, 0.3, 0.5, 0.7, and 0.9, and only the parameters  $t_{epoch}$ ,  $P$ ,  $R_p/R_s$ , and  $a/R_s$  are allowed to vary. After the completion of the reduced-parameter fits, the all-transit fit and the odd-even transit fit follow with the initial values set to the fitted parameters of the reduced-parameter fit with the minimum  $\chi^2$  metric and the corresponding fixed value of the impact parameter.

Figure 12.11 shows the diagnostic plots of the reduced parameter fits of the 6th TCE of the target star KIC 6541920. As shown in the figure, as the fixed value of  $b$  increases from 0.1 to 0.9, the  $\chi^2$  metric varies less than 0.2% in the reduced parameter fits. However,  $R_p/R_s$  increases by around 20% and  $a/R_s$  decreases by more than 50%. The outputs of the reduced parameter fit with the minimum  $\chi^2$  metric are labeled with red dashed lines in the figure. As illustrated in Figure 12.6 of Section 12.4, an increase in  $R_p/R_s$  leads to an increase in the transit depth, an increase in  $a/R_s$  leads to a decrease in the transit duration, and an increase in  $b$  results in the decrease in both the transit depth and duration. The observations of Figure 12.6 are consistent with the systematic variations in  $R_p/R_s$  and  $a/R_s$  versus  $b$  in the reduced parameter fits shown in Figure 12.11: when the fixed value of  $b$  increases, both the transit depth and duration tend to decrease. Therefore,  $R_p/R_s$  increases and  $a/R_s$  decreases to compensate for the effect of the increase of  $b$ , so that a good fit of the model light curve to the flux time series is achieved.

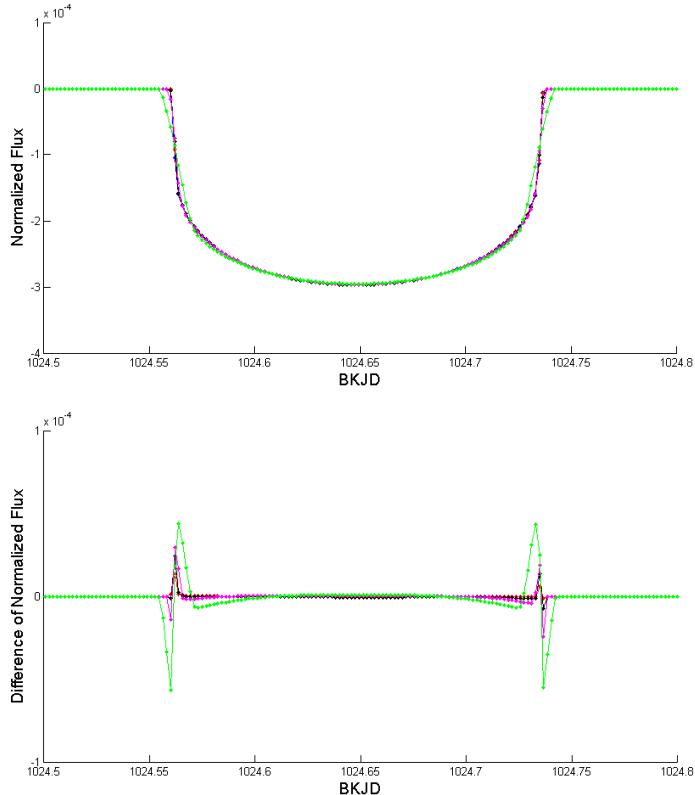


**Figure 12.11** Reduced parameter fits of the 6th TCE of KIC 6541920:  $\chi^2$  metric (top), fitted parameters  $R_p / R_s$  (middle), and  $a / R_s$  (bottom) vs. impact parameter,  $b$ .

The plot on the top of Figure 12.12 shows the light curves generated by the geometric transit signal generator with the fixed values of  $b$  and the corresponding sets of fitted parameters  $t_{epoch}$ ,  $P$ ,  $R_p / R_s$ , and  $a / R_s$  of the reduced parameter fits of the 6th TCE of KIC 6541920, where the light curves corresponding to the fixed  $b$  values of 0.1, 0.3, 0.5, 0.7, and 0.9 are plotted as blue, red, black, magenta, and green lines, respectively. The plot on the bottom of Figure 12.12 shows the differences between light curves with fixed  $b$  values of 0.3, 0.5, 0.7, and 0.9 and the one with  $b = 0.1$ . It is observed that the difference in the light curves with different values of  $b$  is small; therefore, any small variation in the input flux time series may result in a large change in the fitted parameters of  $R_p / R_s$ ,  $a / R_s$ , and  $b$  in the all-transit fit and odd-even transit fit due to accumulated numerical errors, as discussed in the first paragraph of this section.

### 12.5.3 Odd-Even Transit Fit

When the fitting of all transits converges successfully, the same fitting algorithm is executed to fit the odd and even transits to the geometric transit model. The outputs of the odd-even transit

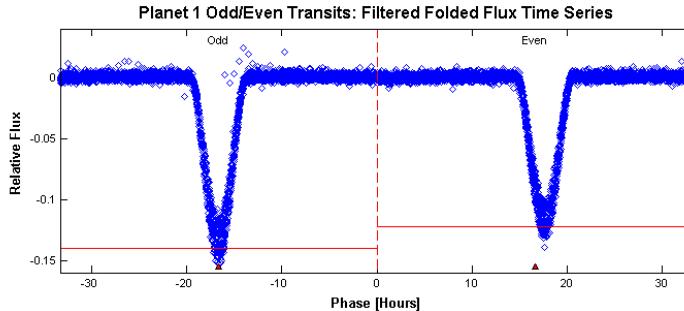


**Figure 12.12** The plot on the top shows light curves generated with the geometric transit signal generator with the fixed  $b$  values of 0.1, 0.3, 0.5, 0.7, and 0.9 and the corresponding sets of fitted parameters  $t_{epoch}$ ,  $P$ ,  $R_p/R_s$ , and  $a/R_s$  of the reduced parameter fits of the 6th TCE of KIC 6541920. The plot on the bottom shows the differences between light curves with fixed  $b$  values of 0.3, 0.5, 0.7, and 0.9 and one with  $b = 0.1$ .

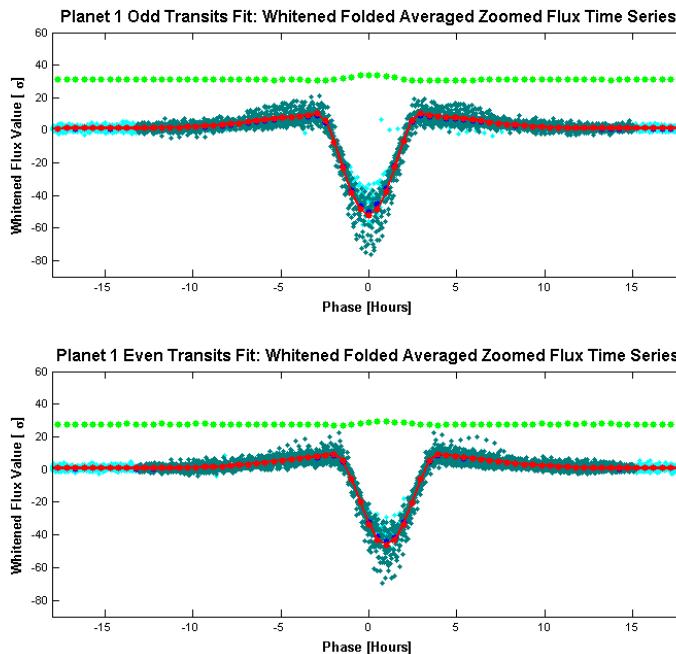
fits are used in the diagnostic tests of the DV module to identify false positives generated by an eclipsing binary target or a target with an eclipsing binary in the background.

The depths of multiple transits of a planet are ideally the same, and the transits of a planet are evenly spaced in time (in the absence of significant transit-timing variations). In contrast, the depths of primary and secondary eclipses of an eclipsing binary system are generally different due to the difference in size and brightness of the two stars. Furthermore, the difference in the epoch times of the primary and secondary eclipses is usually not equal to half of the orbital period of the eclipsing binary system, since the orbit of two stars moving around their gravitational center is generally not circular. In the odd-even transit fit, two sets of parameters, one set for odd transits and the other set for even transits, are determined through an iterative whitening and model fitting process described in Subsection 12.5.1, and the derived parameters are calculated for each. For each TCE, the transit depths and epochs and the corresponding uncertainties derived from the odd-even transit fit are used in the eclipsing binary discrimination tests to distinguish the flux time series of an eclipsing binary system whose primary and secondary eclipses are identified as one TCE in the TPS module. That is, the trial orbital period identified in TPS is half the true orbital period, so that the primary eclipses are folded on top of the secondary eclipses. The details of the eclipsing binary discrimination tests in the DV module are discussed in Chapter 11.

Figure 12.13 shows the folded unwhitened flux time series of the odd and even transits of the eclipsing binary target KIC 6960446. Figure 12.14 shows the folded whitened flux time series of the odd and even transits of the same target and the folded whitened model light curves generated with fitted parameters of the odd and even transits, respectively. As shown in the figures, the primary and secondary eclipses are identified as one TCE by the TPS module, the fits of odd and even transits, which are actually primary and secondary eclipses, demonstrate that the derived transit depths of odd and even transits are different by around 15% and that the transit epoch time of the even transits has a small offset of around an hour relative to that of the odd transits.



**Figure 12.13** Folded unwhitened flux time series of the odd (left) and even (right) transits of KIC 6960446.



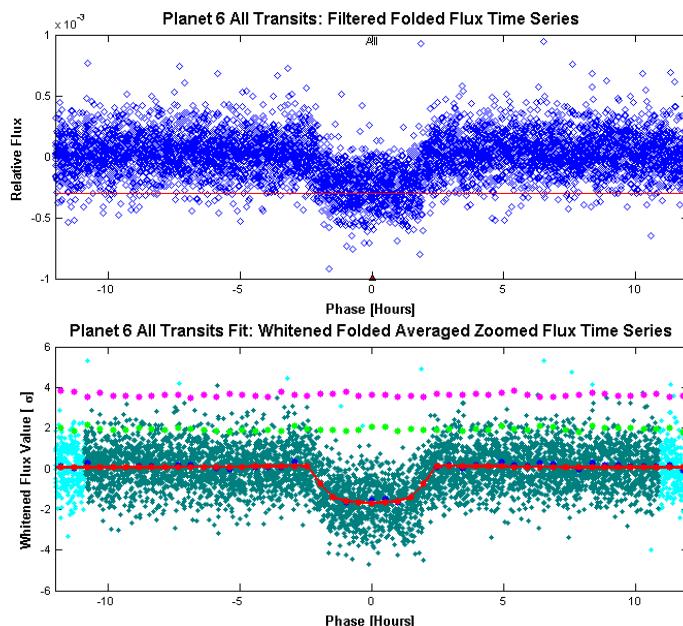
**Figure 12.14** Folded flux time series and model light curves, both in whitened domain, of the odd (top) and even (bottom) transits of KIC 6960446.

### 12.5.4 Outputs of Geometric Transit Model Fits

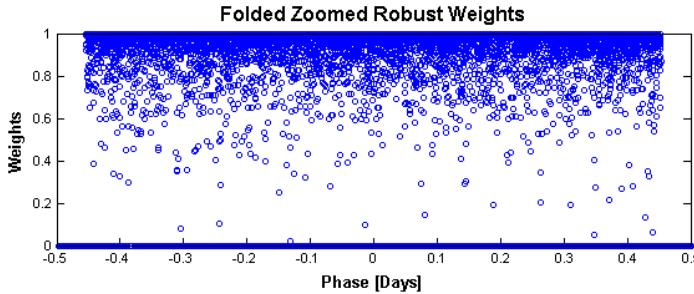
When a TCE is identified in the multiple-planet search, as described later in Section 12.7, a simple check is implemented before the fitting of the TCE. When the eclipsing depth is more than 250,000 ppm, the TCE is labeled as a suspected eclipsing binary and no geometric transit model fitting is performed.

When a geometric transit model fitting algorithm is completed successfully, the fitted parameters and the uncertainties, the derived parameters and the uncertainties, and the goodness of fit metrics, etc. are saved in the DV output structure. In addition, a set of diagnostic plots are generated by the geometric transit model fitting algorithm. As examples, the diagnostic plots of the all-transit fit of the 6th TCE of the target star KIC 6541920 are shown in Figure 12.15, Figure 12.16, and Figure 12.17.

Figure 12.15a shows the folded unwhitened flux time series of all transits of the TCE, and Figure 12.15b shows the corresponding folded whitened flux time series in the same phase range. It is noted that the vertical scales of the two plots are different: the unwhitened flux in Figure 12.15a is dimensionless while the whitened flux in Figure 12.15b is in units of the standard deviation of the unwhitened flux. The transit depth derived from the all-transit fit is illustrated with a horizontal red line in Figure 12.15a. In Figure 12.15b, the folded whitened light curve is illustrated in red, which is generated by the geometric transit signal generator with the fitted parameters derived from the robust fit of all transits. The flux data whose robust weights are larger than 0.1 in the all-transit fit are plotted as dark green dots, otherwise, in light blue dots. The residuals of the fit, determined as the difference of the binned average values of the whitened flux and the whitened model light curve, are plotted as green dots. The same residuals, offset by  $180^\circ$  in phase, are plotted as magenta dots, to aid in the detection of the secondary eclipse. Figure 12.16 shows the folded weights of the robust fit of the all-transit fit of the 6th TCE of the target star KIC 6541920, in the same phase range as Figure 12.15.

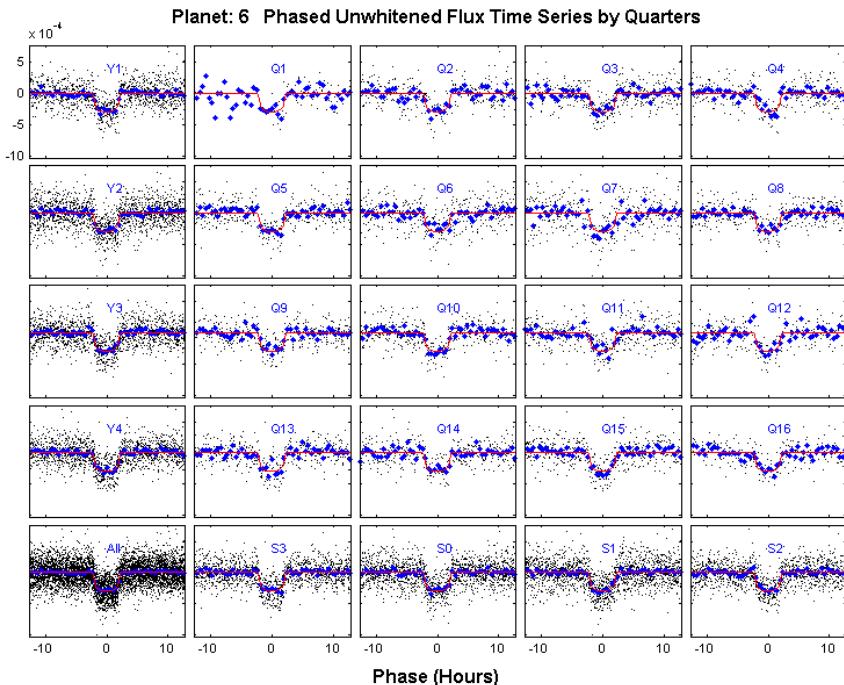


**Figure 12.15** Folded flux time series and model light curve of the all-transit fit of the 6th TCE of KIC 6541920: a) unwhitened flux (top), b) whitened flux and whitened model light curve (bottom).



**Figure 12.16** Folded robust weights of the all-transit fit of the 6th TCE of KIC 6541920.

Figure 12.17 shows the folded unwhitened flux time series of the transits of the 6th TCE of the target star KIC 6541920 by quarter and season, as well as the corresponding folded unwhitened model light curves of the all-transit fit. The folded transits from the same year of the *Kepler* mission are plotted in the same row, and the folded transits in the same season are plotted in the same column. For example, the folded transits in Q4 are shown in the upper right corner of the figure. The folded transits of the first year, including Q1, Q2, Q3, and Q4, are shown in the upper left corner, and the folded transits in Season 2, including Q4, Q8, Q12, and Q16, are shown in the lower right corner. At the lower left corner, the folded transits in all 17 quarters of the *Kepler* science data are illustrated.



**Figure 12.17** Folded flux time series of the transits and folded model light curves of the all-transit fit, both unwhitened, of the 6th TCE of KIC 6541920 by quarter and season.

For the odd-even transit fit, as illustrated in Figure 12.13 and Figure 12.14, the plots of folded unwhitened flux time series of odd and even transits are placed horizontally, and the plots of folded flux time series and folded model light curves, both whitened, are placed vertically, so that the difference in the derived depths and the offset in the fitted transit epoch times of the odd and even transits can be easily observed.

For the reduced parameter fits, a set of diagnostic plots, the same as those of the all-transit fit, are generated for each fit. In addition, as illustrated in Figure 12.11, several diagnostic plots are generated to illustrate the variations of the  $\chi^2$  metric and fitted parameters versus the fixed value of the impact parameter, and the fit with the minimum  $\chi^2$  metric is labeled with red dash lines.

### 12.5.5 Alerts of Failed Fitting Cases

When the geometric transit model fit fails, an alert is generated indicating what the failure is and where it occurs. Table 12.3 lists the top five alerts of the failed all-transit fits.

As shown in the table, the most common failure of the all-transit fits is that the time used by the fitting algorithm goes beyond the preset limit and the fit is stopped during the call of the function “model\_function.” This usually happens when an anomalously noisy flux time series is fitted to a transit model; the criterion of convergence can never be met and the algorithm goes into a never-ending loop.

In the fitting algorithm, several check points are set to verify the validity of the fit results. For example, the fitted parameter of the transit epoch time  $t_{epoch}$  should fall in a range centered at the corresponding TCE value given by the TPS module; the derived transit duration cannot be smaller than the duration of a LC interval (29.4 min) or else the fit results are labeled as invalid. As shown in items 2 and 3 of Table 12.3, the alerts of invalid fit results are generated during the call of the functions “fit\_transit” and “fit\_planet\_results.”

The fourth alert of Table 12.3 is generated when the time used in the iterative numerical integration algorithm, as described in Section 12.4, goes beyond the preset limit in the call of the function to compute the transit light curve when the small-body approximation is not applicable. The fifth alert occurs during the call of the function “transitFitClass” when too many flux data points are gapped and the number of remaining transits is less than 2 in the all-transit fit; as a result, there is insufficient information to determine reliable parameters of the transit model.

**Table 12.3** Top five alerts of failed all-transit fits.

Index	Alert Type	Number
1	dv:modelFunction:fitTimeLimitExceeded	1,012
2	dv:fitTransit:transitEpochBkjdBigDifferenceFromTceValue	592
3	dv:fillPlanetResults:transitDurationSmallerThanLowerBound	262
4	dv:computeLargeBodyTransitLightCurve:takingTooLong	45
5	dv:transitFitClass:insufficientTransitsToFit	41

### 12.6 Trapezoidal Model Fitting Algorithm

As an optional configuration of the transit model fitting in the DV module, the light curve of the target for which a TCE is generated can also be fitted by a trapezoidal model. Since the limb-darkening model is not included, the trapezoidal model is a simple description of the basic characteristics of the transits, and may converge to a successful fit when the limb-darkened transit model fit fails. In these cases, the trapezoidal model fit parameters can be used to enable the diagnostic tests to be conducted, which otherwise could not be performed (see Chapter 11).

The trapezoidal model includes the following four fitted parameters:

- Transit epoch time  $t_{epoch}$  (BKJD): same as the fitted parameter of the geometric model defined in Subsection 12.3.1;
- Transit depth  $D$  (ppm): same as the derived parameter of the geometric model defined in Subsection 12.3.2;
- Transit mid-duration  $d_{mid}$  (hours): the duration of transit at half of the transit depth, as illustrated in Figure 12.4;
- Ratio of ingress time to mid-duration  $d_{in}/d_{mid}$  (dimensionless): the transit ingress time  $d_{in}$  is same as the derived parameter defined in Subsection 12.3.2, but this is the ratio of the ingress time to mid-duration.

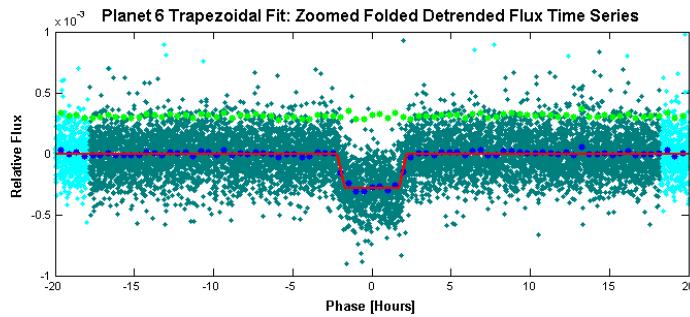
The orbital period  $P$  (days), which is set to the corresponding TCE parameter value provided by the TPS module, is used in the transit signal generator with the trapezoidal model.

An alternative detrending algorithm based on the nonparametric penalized least squares method from Garcia (2010) is applied to the PDC light curve prior to the trapezoidal model fit. The algorithm allows for missing data via weight assignment and solves for the free parameter controlling the amount of smoothing using a generalized cross validation method. To prevent suppression of the transit signal we treat data in transit according to the TCE ephemeris and transit duration as missing with a weight of zero. Each quarter of PDC light curve is detrended independently. When high frequency (similar or shorter time scale than the transit signal) astrophysical signal is present in a light curve, the automated method for determining the smoothing parameter results in unwanted suppression of the transit signal. To prevent over-smoothing, the smoothing parameter is determined on a light curve with a low-pass filter applied. The low-pass filtered light curve is generated by subtraction of a high-pass (simple two-point difference) filtered version of the light curve. The adopted detrending model, which results from using the smoothing parameter estimated from the low-pass filtered version of the light curve, is used in the normalization of the PDC light curve.

The trapezoidal model fitting algorithm is implemented with 10 repeated LM fits. For each fit, the initial value of the fitted parameter is set randomly with a uniform distribution in a pre-determined range. The outputs of the trapezoidal model fitting algorithm are determined as those of the LM fit with the minimum  $\chi^2$  metric.

Figure 12.18 shows a diagnostic plot generated in the trapezoidal model fit of the 6th TCE of KIC 6541920. Only the flux data whose timestamps fall in the time ranges of 8 times of transit duration (one of the TCE parameters generated by the TPS module) and centered at the transit center time are used in the trapezoidal model fit. The flux data points within this range used in the fit are plotted as dark green dots in the figure, otherwise, in light blue dots. The folded light curve generated by the trapezoidal model with the fitted parameters is plotted as red lines and the residual of the fit is plotted as green dots. Since the whitening filter, described in Subsubsection 12.5.1.1, is not used in the trapezoidal model fitting algorithm, all the data shown in Figure 12.18 are in the unwhitened domain. Compared to Figure 12.15b of Subsection 12.5.4, the bottom of the transit is flat in the model light curve shown in Figure 12.18 since the limb-darkening effect is not included in the trapezoidal transit model.

The trapezoidal model fit provides a quick and robust assessment of the transit signal. The fitted trapezoidal transit model is used in the diagnostic tests of the DV module when the fit with the geometric transit model fails or when the fit is not performed, such as for suspected eclipsing binaries.



**Figure 12.18** Folded flux time series and folded model light curve of the trapezoidal model fit, both unwhitened, of the 6th TCE of KIC 6541920.

## 12.7 Multiple-Planet Search

After the fitting process has completed, the transits are removed based on the fitted parameters of the all-transit fit and the residual flux is subjected to a search for additional planets by calling TPS in the DV module. The transit model fitting algorithms, including the reduced parameter fits, all-transit fit, odd-even transit fit, and the trapezoidal model fit, are applied if an additional TCE is generated. The search for additional planets concludes when no additional TCEs are produced or an iteration limit is reached, as shown in the flowchart of Chapter 12.

Figure 12.19 shows the light curve of KIC 6541920 (*Kepler*-11) from Q1 to Q4. The transits of six TCEs are labeled with different colors and symbols in the figure. The first TCE, labeled with red circles, is identified by the TPS module and the corresponding parameters to characterize the TCE included in the DV input structure. The remaining five TCEs are identified in the multiple-planet search by calling TPS in the DV module.

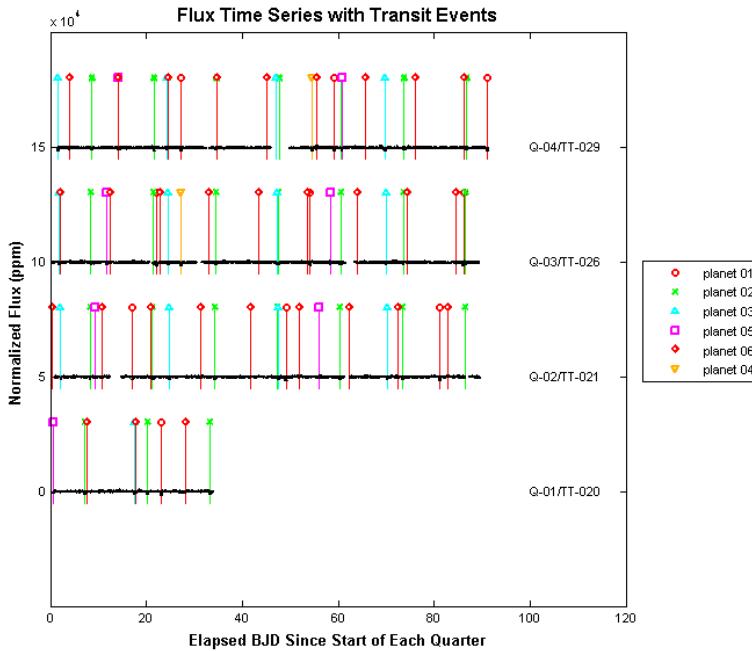
Figure 12.20 shows the folded flux time series of KIC 6541920 in the unwhitened domain, phased with the fitted parameters  $t_{epoch}$  and  $P$  of the 5th and 6th TCEs, respectively. The binned average values of the folded flux and the folded model light curve are plotted as blue and red dots, respectively. The triangles in different colors show the location of the transits of all six TCEs in the phased flux time series.

## 12.8 Performance of Transit Model Fitting

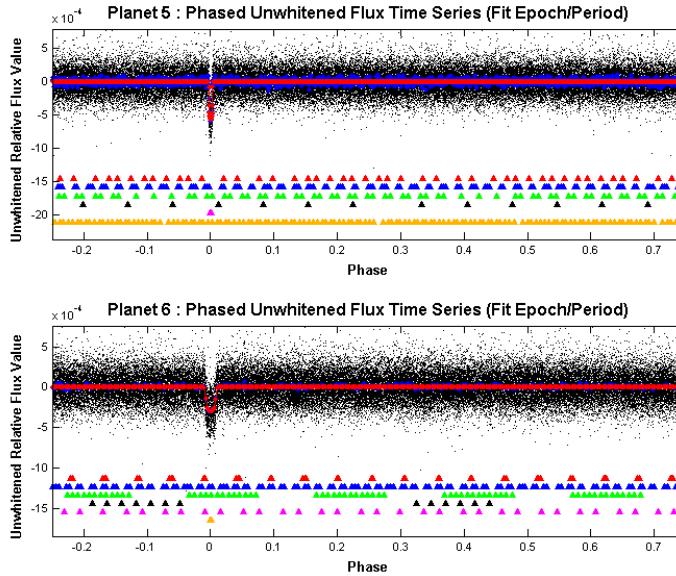
The 17 quarters of science data, collected by the *Kepler* spacecraft from May 13, 2009 to April 8, 2013, were processed by the SOC 9.3 codebase of the *Kepler* Data Processing Pipeline in January 2016. 17,230 target stars, which generated the TCEs in the TPS module, were processed by the DV module successfully.

Among a total of 34,032 TCEs generated in the TPS module and in the multiple-planet search of the DV module, 239 TCEs were labeled as suspected eclipsing binaries, 2062 TCEs failed in the all-transit fit, and 31,731 TCEs completed the all-transit fit successfully. Out of 31,731 TCEs with successful all-transit fits, 2620 TCEs failed in the odd-even transit fit, and 29,111 TCEs completed the odd-even transit fit successfully. 33,125 out of 34,032 TCEs completed the trapezoidal model fit successfully.

Figure 12.21 compares the orbital period of the DV all-transit fit and the corresponding KOI parameter. The plot on the left shows all orbital periods in the comparison and the plot on the right shows the orbital periods ranging from 0 to 20 days only. The diagonal green line shows the DV fitted orbital period value is equal to the KOI parameter value; the other four green lines indicate that the two period values differ by a factor of 1/3, 1/2, 2, and 3, respectively. It is



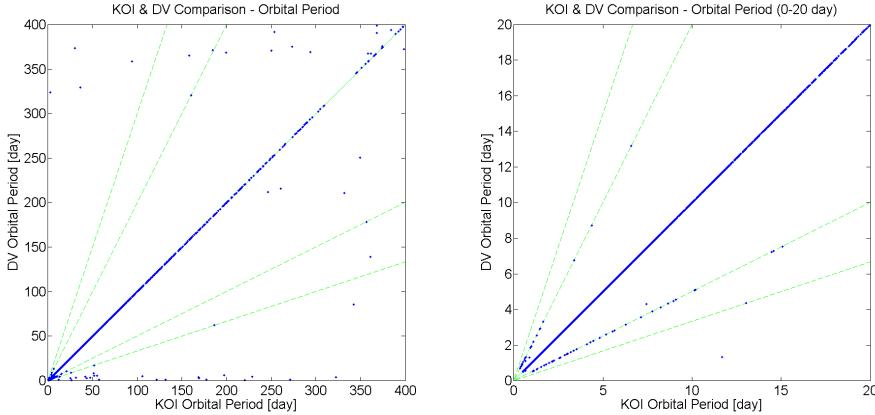
**Figure 12.19** Light curve of KIC 6541920 from Q1 to Q4 and transits of six TCEs.



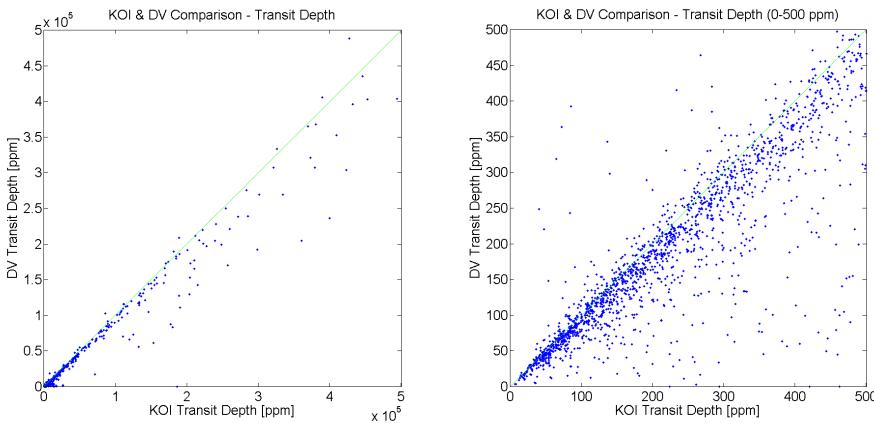
**Figure 12.20** Phased flux time series of KIC 6541920 with the fitted parameters  $t_{epoch}$  and  $P$  of the 5th (top) and 6th (bottom) TCEs, respectively.

observed from Figure 12.21 that the orbital periods of some TCEs identified in TPS and DV are double or half of the corresponding KOI values.

Figure 12.22 compares the transit depth derived from the DV all-transit fit and the corresponding KOI parameter. Similar to Figure 12.21, the plot on the left shows all-transit depths in the comparison and the plot on the right shows the transit depths ranging from 0 to 500 ppm only. The diagonal green line shows that the DV fitted transit depth value is equal to the KOI parameter value. It is observed the KOI values of the transit depth are larger than the corresponding DV fitted values for some TCEs. Investigations show some transit signals are reduced in the light curve preprocessing procedure of harmonic removal when the orbital period is small.



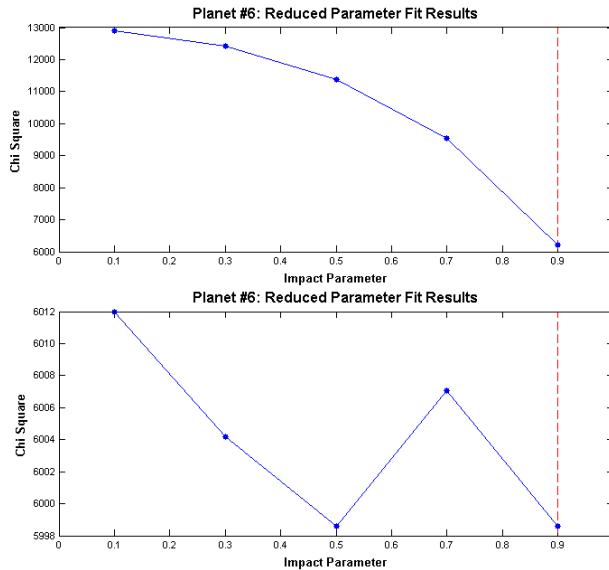
**Figure 12.21** Comparison of DV Fitted parameters and KOI parameters: all orbital periods (left) and orbital periods ranging from 0 to 20 days (right).



**Figure 12.22** Comparison of DV Fitted parameters and KOI parameters: all-transit depths (left) and transit depths ranging from 0 to 500 ppm (right).

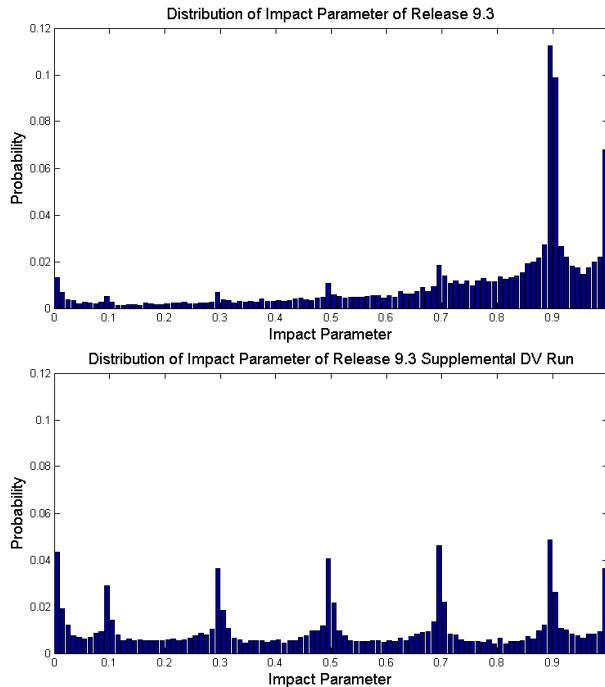
A software defect introduced into the SOC 9.3 code for the reduced parameter fits came to light after its release. As discussed in Subsection 12.5.1, only the data points within the range of the transit and a buffer on each side of the transit are used in the weighted nonlinear least-squares fitting, and the weights are assigned 1 and 0, respectively, depending on whether the data points are used in the fitting or not. As shown in Equation 12.12, the  $\chi^2$  metric is related to how many data points are used in the fit: the more data points used in the fit, the larger the  $\chi^2$  metric. In the SOC 9.3 codebase, the data points used in the reduced parameter fits are related to the fixed

value of the impact parameter  $b$ . As a result, the calculated  $\chi^2$  metric is improperly related to the value of  $b$ : the closer  $b$  is to 1, the smaller the  $\chi^2$  metric. The software defect was corrected in the modified SOC 9.3 codebase, which was used in the supplemental DV run in August 2016. Figure 12.23 shows the diagnostic plots of the  $\chi^2$  metric versus  $b$  of the reduced parameter fits of the 6th TCE of KIC 6541920, which were generated by the SOC 9.3 codebase in January 2016 and the modified SOC 9.3 codebase in August 2016, respectively. As shown in the plot on the top of Figure 12.23, due to the software defect, the  $\chi^2$  metric systematically decreases as  $b$  increases, so the outputs of the reduced parameter fit with the fixed value of  $b = 0.9$  are always selected to seed the all-transit fit. In the plot on the bottom of Figure 12.23, there is no systematic decrease of the  $\chi^2$  metric as  $b$  increases.



**Figure 12.23** The diagnostic plots of  $\chi^2$  versus  $b$  of the reduced parameter fits of the 6th TCE of KIC 6541920, generated by the SOC 9.3 codebase in January 2016 (top) and the modified SOC 9.3 codebase in August 2016 (bottom), respectively. As shown in the plot on the top, due to a software defect introduced into the 9.3 codebase, the  $\chi^2$  metric of the reduced parameter fit systematically decreases as the fixed value of the impact parameter  $b$  increases. In the plot on the bottom, there is no systematic decrease of the  $\chi^2$  metric as  $b$  increases.

As shown in Subsection 12.5.2, the flux time series with a low SNR or of transiting planet signatures for small planets (relative to the size of their host stars) may be well fitted over a wide range of impact parameter values. Figure 12.24 shows the distributions of the fitted parameter  $b$  in the all-transit fits of a set of 16,514 TCEs with the SOC 9.3 codebases in January 2016 and the modified SOC 9.3 codebase in August 2016, respectively. It is observed that the convergence of the all-transit fit is essentially independent of the initial seed value of the impact parameter  $b$  for the large planets; however, the distribution of the fitted parameter  $b$  for the small planets is biased toward the initial seed value of  $b = 0.9$  in the outputs of the all-transit fits with the SOC 9.3 codebase, as shown in the plot on the top of Figure 12.24. In the plot on the bottom of Figure 12.24, there is no bias toward  $b = 0.9$  in the distribution of the fitted parameter  $b$  in the all-transit fits with the modified SOC 9.3 codebase. As discussed in detail by Twicken et al. (2016), transiting planets with a high impact parameter must be larger than those with a lower impact parameter for given transit depths on the same host stars because of the limb-darkening effect. It is noted that all planetary candidates in the DR25 *Kepler* Mission catalog will be modeled independently by the TCE Review Team (TCERT), so the bias discussed here relates only to



**Figure 12.24** Distribution of the fitted parameter  $b$  of the all-transit fits of a set of 16,514 TCEs, generated by the SOC 9.3 codebase in January 2016 (top) and the modified SOC 9.3 codebase in August 2016 (bottom), respectively. As shown in the plot on the top, the distribution of the fitted parameter  $b$  of the small planets is biased toward  $b = 0.9$  in the outputs of the all-transit fit of the SOC 9.3 codebase. In the plot on the bottom, there is no bias toward  $b = 0.9$  in the distribution of the fitted parameter  $b$  in the outputs of the all-transit fit of the modified SOC 9.3 codebase.

TCE products of the SOC 9.3 Data Release 25 of the *Kepler* Science Data Processing Pipeline at the NASA Exoplanet Archive.

## 12.9 Conclusions

We have presented the algorithms of the transit model fitting of the Data Validation module of the *Kepler* Science Data Processing Pipeline. The performance of the algorithms is demonstrated by the outputs of processing 17 quarters of *Kepler* science data using SOC 9.3 of the *Kepler* Science Data Processing Pipeline in January 2016. The results of the transit model fitting of the TCEs identified by the pipeline are accessible by the science community at NASA Exoplanet Archive.

## Bibliography

- Claret, A., & Bloemen, S., 2011. “Gravity and Limb-Darkening Coefficients for the *Kepler*, CoRoT, Spitzer, uvby, UVRIJHK, and Sloan photometric systems,” *Astronomy & Astrophysics*, 529, A75, available online at <http://dx.doi.org/10.1051/0004-6361/201116451>
- Garcia, D., 2010. “Robust Smoothing of Gridded Data in One and Higher Dimensions with Missing Values,” *Computational Statistics and Data Analysis*, 54, 1167

- Jenkins, J. M., 2002. "The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets," *ApJ*, 575, 493
- Mandel, K., & Agol, E., 2002. "Analytic Light Curves for Planetary Transit Searches," *ApJL*, 580, L171
- Tenenbaum, P., Bryson, S. T., Chandrasekaran, H., et al. 2010. "An Algorithm for the Fitting of Planet Models to Kepler Light Curves," in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400J
- Twicken, J. D., Jenkins, J. M., Seader, S. E., et al., 2016. "Detection of Potential Transit Signals in 17 Quarters of Kepler Data: Results of the Final Kepler Mission Transiting Planet Search (DR25)," *AJ*, 152, 158
- Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010. "Data Validation in the Kepler Science Operations Center Pipeline," in Proc. SPIE, Vol. 7740, 42W



## CHAPTER 13

---

### ACRONYMS AND ABBREVIATION LIST

**ADB** Array Data Base

**ADC** Analog-to-Digital Converter

**ADU** Analog-to-Digital Unit

**AED** Ancillary Engineering Data

**AR** Archive to DMC Pipeline Module

**ARP** Artifact Removal Pixels

**AU** Astronomical Unit

**BART** 2-D Black and Artifact Removal Tool

**BDT** Barycentric Dynamic Time

**BKJD** Barycentric-corrected Kepler Julian Date

**BMJD** Barycentric-corrected Modified Julian Date

**CAL** Calibration Pipeline Module

**CCD** Charge Coupled Device

**CDF** Cumulative Distribution Function

**CDPP** Combined Differential Photometric Precision

**CDQ** Check Data Quality Pipeline Module

**CM** Catalog Management Pipeline Module

**COA** Create Optimal Apertures Pipeline Submodule in TAD

**COMP** COMPression Pipeline Module

**CSCI** Computer Software Configuration Item

**CTE** Charge Transfer Efficiency

**DAWG** Data Analysis Working Group

**DG** Data Goodness Pipeline Module

**DMC** Data Management Center

- DR** Data Receipt Pipeline Module
- DS** Data Store Pipeline Module
- DSN** Deep Space Network
- DV** Data Validation Pipeline Module
- DVA** Differential Velocity Aberration
- DYN** Dynamic Black (aka Dynablack) Pipeline Module
- Dec** Declination
- EPIC** Ecliptic Plane Input Catalog
- ETEM** End-To-End Model Pipeline Module
- FAP** False Alarm Probability
- FAR** False Alarm Rate
- FC** Focal plane Characterization
- FFI** Full Frame Image
- FFT** Fast Fourier Transform
- FGS** Fine Guidance Sensor
- FITS** Flexible Image Transport System
- FOP** Follow-up Observation Program
- FOV** Field of View
- FP** False Positive
- FPG** Focal Plane Geometry model
- GOF** Goodness of Fit metric
- GO** Guest Observer
- HAC** Histogram Accumulator in COMP
- HAG** Histogram Aggregator in COMP
- HGA** High Gain Antenna
- HGN** Histogram GeNerator in COMP
- HZ** Habitable Zone
- KADN** Kepler Ames Design Notes
- KIC** Kepler Input Catalog
- KJD** Kepler-modified Julian Date
- KOI** Kepler Object of Interest

**LC** Long Cadence

**LDE** Local Detector Electronics

**LM** Levenberg-Marquardt algorithm

**LS** Least-Squares

**LSI** Linear Shift-Invariant system

**LUN** Logical Unit Number

**MAD** Median Absolute Deviation

**MAP** Maximum A Posteriori

**MAST** Mikulski Archive for Space Telescopes

**MES** maximum Multiple Event Statistic

**MJD** Modified Julian Date

**MLE** Maximum Likelihood Estimator

**MLM** Maximum Likelihood Method

**MMO** Mission Management Office

**MMOC** Multi-Mission Operations Center

**MOC** Mission Operations Center

**MOM** Message Oriented Middleware

**MPD** Moiré Pattern

**MR** Mission Reports Pipeline Module

**NAS** NASA Advanced Supercomputing Division

**NExSci** NASA Exoplanet Science Institute

**OWT** Overcomplete Wavelet Transform

**PA** Photometric Analysis Pipeline Module

**PAD** PPA Attitude Determination Pipeline Module

**PAG** PMD Aggregator Pipeline Module

**PC** Planet Candidate

**PCHIP** Piecewise Cubic Hermite Interpolation

**PDC** Pre-Search Data Conditioning Pipeline Module

**PDC-MAP** Pre-Search Data Conditioning Maximum A Posteriori algorithm

**PDC-msMAP** Pre-Search Data Conditioning MultiScale Maximum A Posteriori algorithm

**PDF** Probability Density Function

- PDQ** Photometer Data Quality Pipeline Module
- PI** Pipeline Infrastructure Pipeline Module
- PMD** PPA Metrics Determination
- POOF** Pixel Overlay On FFI
- POU** Propagation of Uncertainties
- PPA** Photometer Performance Assessment Pipeline Module
- ppm** Parts-per-million
- PRF** Pixel Response Function
- PSD** Power Spectral Density
- PSF** Point Spread Function
- RMOM** Remote Message Oriented Middleware
- rms** Root Mean Square
- ROI**
- RP** Reference Pixel
- RS**
- SAN** Storage Area Network
- SAP** Simple Aperture Photometry
- SAS** Science Analysis System
- SBT** Sandbox Tools
- SC** Short Cadence
- SDF** Suspect Data Flag
- SES** Single Event Statistic
- SLIB** Support LIBraries
- SNR** Signal-to-Noise Ratio
- ssh** Secure Shell Protocol
- SO** Science Office
- SOC** Science Operations Center
- SOL** Start of Line
- SOLR** Start of Line Ringing
- SPSD** Sudden Pixel Sensitivity Dropout
- SPWG** Stellar Properties Working Group

**SSR** Solid State Recorder

**STFT** Short Time Fourier Transform

**STScI** Space Telescope Science Institute

**SVD** Singular Value Decomposition

**TAD** Target and Aperture Definitions Pipeline Module

**TCAT** Temperature Coefficient Analysis Tool

**TCE** Threshold Crossing Events

**TCERT** Threshold Crossing Event Review Team

**TESS** Transiting Exoplanet Survey Satellite

**TPS** Transiting Planet Search Pipeline Module

**TTV** Transit Timing Variation

**TUS** Target Under Study

**UOW** Unit Of Work

**UTC** Coordinated Universal Time

**WGN** White Gaussian Noise

**ZMUV** Zero-Mean, Unit-Variance WGN

