Draft version October 13, 2017
Typeset using LaTeX **twocolumn** style in AASTeX61

PLANETARY CANDIDATES OBSERVED BY *Kepler*. VIII. A FULLY AUTOMATED CATALOG WITH MEASURED COMPLETENESS AND RELIABILITY BASED ON DATA RELEASE 25

Susan E. Thompson,[1,2,3,*] Jeffrey L. Coughlin,[2,1] Kelsey Hoffman,[1] Fergal Mullally,[1,2,4] Jessie L. Christiansen,[5] Christopher J. Burke,[2,1,6] Steve Bryson,[2] Natalie Batalha,[2] Michael R. Haas,[2,†] Joseph Catanzarite,[1,2] Jason F. Rowe,[7] Geert Barentsen,[8] Douglas A. Caldwell,[1,2] Bruce D. Clarke,[1,2] Jon M. Jenkins,[2] Jie Li,[1] David W. Latham,[9] Jack J. Lissauer,[2] Savita Mathur,[10] Robert L. Morris,[1,2] Shawn E. Seader,[11] Jeffrey C. Smith,[1,2] Todd C. Klaus,[2] Joseph D. Twicken,[1,2] Jeffrey E. Van Cleve,[1] Bill Wohler,[1,2] Rachel Akeson,[5] David R. Ciardi,[5] William D. Cochran,[12] Christopher E. Henze,[2] Steve B. Howell,[2] Daniel Huber,[13,14,1,15] Andrej Prša,[16] Solange V. Ramírez,[5] Timothy D. Morton,[17] Thomas Barclay,[18] Jennifer R. Campbell,[2,19] William J. Chaplin,[20,15] David Charbonneau,[9] Jørgen Christensen-Dalsgaard,[15] Jessie L. Dotson,[2] Laurance Doyle,[21,1] Edward W. Dunham,[22] Andrea K. Dupree,[9] Eric B. Ford,[23,24,25,26] John C. Geary,[9] Forrest R. Girouard,[27,2] Howard Isaacson,[28] Hans Kjeldsen,[15] Elisa V. Quintana,[18] Darin Ragozzine,[29] Avi Shporer,[30] Victor Silva Aguirre,[15] Jason H. Steffen,[31] Martin Still,[8] Peter Tenenbaum,[1,2] William F. Welsh,[32] Angie Wolfgang,[23,24,‡] Khadeejah A Zamudio,[2,19] David G. Koch,[2,§] and William J. Borucki[2,†]

[1] *SETI Institute, 189 Bernardo Ave, Suite 200, Mountain View, CA 94043, USA*
[2] *NASA Ames Research Center, Moffett Field, CA 94035, USA*
[3] *Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218*
[4] *Orbital Insight, 100 W Evelyn Ave #110, Mountain View, CA 94041*
[5] *IPAC-NExScI, Mail Code 100-22, Caltech, 1200 E. California Blvd. Pasadena, CA 91125*
[6] *MIT Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Avenue, 37-241, Cambridge, MA 02139*
[7] *Dept. of Physics and Astronomy, Bishop's University, 2600 College St., Sherbrooke, QC, J1M 1Z7, Canada*
[8] *Bay Area Environmental Research Institute, 625 2nd St., Ste 209, Petaluma, CA 94952, USA*
[9] *Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge MA 02138, USA*
[10] *Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA*
[11] *Rincon Research Corporation,101 N Wilmot Rd, Tucson, AZ 85711*
[12] *McDonald Observatory and Department of Astronomy, University of Texas at Austin, Austin, TX 78712*
[13] *Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*
[14] *Sydney Institute for Astronomy (SIfA), School of Physics, University of Sydney, NSW 2006, Australia*
[15] *Stellar Astrophysics Centre, Dept. of Physics and Astronomy, Aarhus University, Ny Munkegade 120, 8000 Aarhus C, Denmark*
[16] *Villanova University, Dept. of Astrophysics and Planetary Science, 800 Lancaster Ave, Villanova PA 19085*
[17] *Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA*
[18] *NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771*
[19] *KRBwyle, 2400 Nasa Parkway, Houston, TX 77058 USA*
[20] *School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[21] *Institute for the Metaphysics of Physics, Principia College, One Maybeck Place, Elsah, Illinois 62028*
[22] *Lowell Observatory, 1400 W Mars Hill Rd, Flagstaff, AZ 86001*
[23] *Dept. of Astronomy & Astrophysics, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA, 16802, USA*
[24] *Center for Exoplanets and Habitable Worlds, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA, 16802, USA*
[25] *Center for Astrostatistics, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA, 16802, USA*
[26] *Institute for CyberScience, The Pennsylvania State University*
[27] *Orbital Sciences Corporation, 2401 East El Segundo Boulevard, Suite 200, El Segundo, CA 90245, USA*
[28] *Dept. of Astronomy, UC Berkeley, Berkeley, CA 94720, USA*
[29] *Brigham Young University, Department of Physics and Astronomy, N283 ESC, Provo, UT 84602, USA*
[30] *Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA*

[31] *University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154*
[32] *Department of Astronomy, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-1221*

ABSTRACT

We present the Kepler Object of Interest (KOI) catalog of transiting exoplanets based on searching four years of *Kepler* time series photometry (Data Release 25, Q1–Q17). The catalog contains 8054 KOIs of which 4034 are planet candidates with periods between 0.25 and 632 days. Of these candidates, 219 are new in this catalog and include two new candidates in multi-planet systems (KOI-82.06 and KOI-2926.05), and ten new high-reliability, terrestrial-size, habitable zone candidates. This catalog was created using a tool called the Robovetter which automatically vets the DR25 Threshold Crossing Events (TCEs) found by the *Kepler* Pipeline (Twicken et al. 2016). Because of this automation, we were also able to vet simulated data sets and therefore measure how well the Robovetter separates those TCEs caused by noise from those caused by low signal-to-noise transits. Because of these measurements we fully expect that this catalog can be used to accurately calculate the frequency of planets out to *Kepler*'s detection limit, which includes temperate, super-Earth size planets around GK dwarf stars in our Galaxy. This paper discusses the Robovetter and the metrics it uses to decide which TCEs are called planet candidates in the DR25 KOI catalog. We also discuss the simulated transits, simulated systematic noise, and simulated astrophysical false positives created in order to characterize the properties of the final catalog. For orbital periods less than 100 d the Robovetter completeness (the fraction of simulated transits that are determined to be planet candidates) across all observed stars is greater than 85%. For the same period range, the catalog reliability (the fraction of candidates that are not due to instrumental or stellar noise) is greater than 98%. However, for low signal-to-noise candidates found between 200 and 500 days, our measurements indicate that the Robovetter is 73.5% complete and 37.2% reliable across all searched stars (or 76.7% complete and 50.5% reliable when considering just the FGK dwarf stars). We describe how the measured completeness and reliability varies with period, signal-to-noise, number of transits, and stellar type. Also, we discuss a value called the disposition score which provides an easy way to select a more reliable, albeit less complete, sample of candidates. The entire KOI catalog, the transit fits using Markov chain Monte Carlo methods, and all of the simulated data used to characterize this catalog are available at the NASA Exoplanet Archive.

Keywords: catalogs — planetary systems — planets and satellites: detection — stars: statistics — surveys — techniques: photometric

* a.k.a. Susan E. Mullally, email: smullally@stsci.edu
† NASA Ames Associate
‡ NSF Astronomy & Astrophysics Postdoctoral Fellow
§ deceased

1. INTRODUCTION

Kepler's mission to measure the frequency of Earth-size planets in the Galaxy is an important step towards understanding the Earth's place in the Universe. Launched in 2009, the *Kepler* Mission (Koch et al. 2010; Borucki 2016) stared almost continuously at a single field for four years (or 17, ≈90 day quarters), recording the brightness of ≈200,000 stars (≈160,000 stars at a time) at a cadence of 29.4 minutes over the course of the mission. *Kepler* detected transiting planets by observing the periodic decrease in the observed brightness of a star when an orbiting planet crossed the line of sight from the telescope to the star. *Kepler*'s prime-mission observations concluded in 2013 when it lost a second of four reaction wheels, three of which were required to maintain the stable pointing. From the ashes of *Kepler* rose the *K2* mission which continues to find exoplanets in addition to a whole host of astrophysics enabled by its observations of fields in the ecliptic (Howell et al. 2014; Van Cleve et al. 2016b). While not the first to obtain high-precision, long-baseline photometry to look for transiting exoplanets (see e.g., Barge et al. 2008; O'Donovan et al. 2006), *Kepler* and its plethora of planet candidates revolutionized exoplanet science. The large number of Kepler planet detections from the same telescope opened the door for occurrence rate studies and has enabled some of the first measurements of the frequency of planets similar to the Earth in our Galaxy. To further enable those types of studies, we present here the planet catalog that resulted from the final search of the Data Release 25 (DR25) *Kepler* mission data along with the tools provided to understand the biases inherent in the search and vetting done to create that catalog.

First, we put this work in context by reviewing some of the scientific achievements accomplished using *Kepler* data. Prior to *Kepler*, most exoplanets were discovered by radial velocity methods (e.g. Mayor & Queloz 1995), which largely resulted in the detection of Neptune- to Jupiter-mass planets in orbital periods of days to months. The high precision photometry and the four-year baseline of the *Kepler* data extended the landscape of known exoplanets. To highlight a few examples, Barclay et al. (2013) found evidence for a moon-size terrestrial planet in a 13.3 day period orbit, Quintana et al. (2014) found evidence of an Earth-size exoplanet in the habitable zone of the M dwarf Kepler-186, and Jenkins et al. (2015) statistically validated a super-Earth in the habitable zone of a G-dwarf star. Additionally, for several massive planets *Kepler* data has enabled measurements of planetary mass and atmospheric properties by using the photometric variability along the entire orbit (Shporer et al. 2011; Mazeh et al. 2012; Shporer 2017). *Kepler* data has also revealed hundreds of compact, co-planar, multi-planet systems, e.g., the six planets around Kepler-11 (Lissauer et al. 2011a), which collectively have told us a great deal about the architecture of planetary systems (Lissauer et al. 2011b; Fabrycky et al. 2014). Exoplanets have even been found orbiting binary stars, e.g., Kepler-16 (AB) b (Doyle et al. 2011).

Other authors have taken advantage of the long time series, near-continuous data set of 206,150[1] stars to advance our understanding of stellar physics through the use of asteroseismology. Of particular interest to this catalog is the improvement in the determination of stellar radius (e.g., Huber et al. 2014; Mathur et al. 2017) which can be one of the most important sources of error when calculating planetary radii. *Kepler* data was also used to track the evoluation of star-spots created from magnetic activity and thus enabled the measurement of stellar rotation rates (e.g. Aigrain et al. 2015; García et al. 2014; McQuillan et al. 2014; Zimmerman et al. 2017). Studying stars in clusters enabled Meibom et al. (2011) to map out the evolution of stellar rotation as stars age. *Kepler* also produced light curves of 2876[2] eclipsing binary stars (Prša et al. 2011; Kirk et al. 2016) including unusual binary systems, such as the eccentric, tidally-distorted, Heartbeat stars (Welsh et al. 2011; Thompson et al. 2012; Shporer et al. 2016) that have opened the doors to understanding the impact of tidal forces on stellar pulsations and evolution (e.g., Hambleton et al. 2017; Fuller et al. 2017).

The wealth of astrophysics, and the size of the *Kepler* community, is in part due to the rapid release of *Kepler* data to the NASA Archives: the Exoplanet Archive (Akeson et al. 2013) and the MAST (Mikulski Archives for Space Telescopes). The *Kepler* mission released data from every step of the processing (Thompson et al. 2016a; Stumpe et al. 2014; Bryson et al. 2010), including its planet searches. The results of both the original searches for periodic signals (known as the TCEs or Threshold Crossing Events) and the well-vetted KOIs (Kepler Objects of Interest) were made available for the community. The combined list of *Kepler*'s planet candidates found from all searches can be found in the cumulative KOI table[3]. The KOI table we present here is from a single search of the DR25 light curves. While the search does not include new observations, it was performed using an improved version of the *Kepler* Pipeline (version 9.3, Jenkins 2017a). For a high-level summary of the changes to the *Kepler* Pipeline, see the DR25 data release notes (Thompson et al. 2016b; Van Cleve et al. 2016a). The *Kepler* Pipeline has undergone successive improvements since launch as the data characteristics have become better understood.

[1] This tally only includes the targeted stars and not those observed by "accident" in the larger apertures.

[2] This represents the number reported in the Kepler Binary Catalog, http://keplerebs.villanova.edu, in August 2017.

[3] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=cumulative

The photometric noise at time scales of the transit is what limits *Kepler* from finding small terrestrial-size planets. Investigations of the noise properties of *Kepler* exoplanet hosts by Howell et al. (2016) showed that those exoplanets with the radii $\leq 1.2 R_\oplus$ are only found around the brightest, most photometrically quiet stars. As a result, the search for the truly Earth-size planets are limited to a small subset of *Kepler*'s stellar sample. Analyses by Gilliland et al. (2011, 2015) show that the primary source of the observed noise was indeed inherent to the stars, with a smaller contributions coming from imperfections in the instruments and software. Unfortunately, the typical noise level for 12^{th} magnitude solar-type stars is closer to 30 ppm (Gilliland et al. 2015) than the 20 ppm expected prior to launch (Jenkins et al. 2002), causing *Kepler* to need a longer baseline to find a significant number of Earth-like planets around Sun-like stars. Ultimately, this higher noise level impacts *Kepler*'s planet yield. And, because different stars have different levels of noise, the transit depth to which the search is sensitive varies across the sample of stars. This bias must be accounted for when calculating occurrence rates, and is explored in-depth for this run of the *Kepler* Pipeline by the transit injection and recovery studies of Burke & Catanzarite (2017a,b) and Christiansen (2017).

To confirm the validity and further characterize identified planet candidates, the *Kepler* mission benefited from an active, funded, follow-up observing program. This program used ground-based radial velocity measurements to determine the mass of exoplanets (e.g., Marcy et al. 2014) when possible and also ruled out other astrophysical phenomena, like background eclipsing binaries, that can mimic a transit signal. The follow-up program obtained high-resolution imaging of $\approx$90% of known KOIs (e.g., Furlan et al. 2017) to identify close companions (bound or unbound) that would be included in *Kepler*'s rather large $3.98''$ pixels. The extra light from these companions must be accounted for when determining the depth of the transit and the radii of the exoplanet. While the *Kepler* Pipeline accounts for the stray light from stars in the Kepler Input Catalog (Brown et al. 2011; and see flux fraction in §2.3.1.2 of the Kepler Archive Manual; Thompson et al. 2016a), the sources identified by these high-resolution imaging catalogs were not included. Based on the analysis by Ciardi et al. (2015), where they considered the effects of multiplicity, planet radii are underestimated by a factor averaging $\simeq$1.5 for G dwarfs prior to vetting, or averaging $\simeq$1.2 for KOIs that have been vetted with high-resolution imaging and Doppler spectroscopy. The effect of unrecognized dilution decreases for planets orbiting the K and M dwarfs, because they have a smaller range of possible stellar companions

Even with rigorous vetting and follow-up observations, most planet candidates in the KOI catalogs cannot be directly confirmed as planetary. The stars are too dim and the planets are too small to be able to measure a radial velocity signature for the planet. Statistical methods study the likelihood that the observed transit could be caused by other astrophysical scenarios and have succeeded in validating thousands of *Kepler* planets (e.g. Morton et al. 2016; Torres et al. 2015; Rowe et al. 2014; Lissauer et al. 2014).

The Q1–Q16 KOI catalog (Mullally et al. 2015) was the first with a long enough baseline to be significantly impacted by another source of false positives, the long-period false positives created by the instrument itself. In that catalog (and again in this one), the majority of long-period, low SNR TCEs are ascribed to instrumental effects incompletely removed from the data before the TCE search. *Kepler* has a variety of short timescale (on the order of a day or less), non-Gaussian noise sources including focus changes due to thermal variations, signals imprinted on the data by the detector electronics, noise caused by solar flares, and the pixel sensitivity changing after the impact of a high energy particle (known as a sudden pixel sensitivity drop-out, or SPSD). Because the large number of TCEs associated with these types of errors, and because the catalog was generated to be intentionally inclusive (i.e. high completeness), many of the long-period candidates in the Q1–Q16 KOI catalog are expected to simply be noise. We were faced with a similar problem for the DR25 catalog and spent considerable effort writing software to identify these types of false positives, and for the first time we include an estimate for how often these signals contaminate the catalog.

The planet candidates found in *Kepler* data have been used extensively to understand the frequency of different types of planets in the Galaxy. Many studies have shown that small planets ($< 4 R_\oplus$) in short period orbits are common, with occurrence rates steadily increasing with decreasing radii (Burke & Seader 2016; Howard et al. 2012; Petigura et al. 2013; Youdin 2011). Dressing & Charbonneau (2013, 2015), using their own search, confined their analysis to M dwarfs and orbital periods less than 50 d and determined that multi-planet systems are common around these low mass stars. Therefore planets are more common than stars in the Galaxy (due, in part, to the fact that low mass stars are the most common stellar type). Fulton et al. (2017), using improved measurements of the stellar properties (Petigura et al. 2017a), looked at small planets with periods of less than 100 d and showed that there is a valley in the occurrence of planets near $1.75 R_\oplus$. This result improved upon the results of Howard et al. (2012) and Lundkvist et al. (2016) and further verified the evaporation valley predicted by Owen & Wu (2013) and Lopez & Fortney (2013) for close-in planets.

Less is known about the occurrence of planets in longer period orbits. Using planet candidates discovered with *Kepler*, several papers have measured the frequency of small planets in the habitable zone of sun-like stars (see e.g. Burke et al. 2015; Foreman-Mackey

et al. 2016; Petigura et al. 2013) using various methods. Burke et al. (2015) used the Q1–Q16 KOI catalog (Mullally et al. 2015) and looked at G and K stars and concluded that 10% (with an allowed range of 1–200%) of solar-type stars host planets with radii and orbital periods within 20% of that of the Earth. Burke et al. (2015) considered various systematic effects and showed that they dominate the uncertainties and concluded that improved measurements of the stellar properties, the detection efficiency of the search, and the reliability of the catalog will have the most impact in narrowing the uncertainties in such studies.

1.1. *Design Philosophy of the DR25 catalog*

The DR25 KOI catalog is designed to support rigorous occurrence rate studies. To do that well, it was critical that we not only identify the exoplanet transit signals in the data but also measure the catalog reliability (the fraction of transiting candidates that are not caused by noise), and the completeness of the catalog (the fraction of true transiting planets detected).

The measurement of the catalog completeness has been split into two parts: the completeness of the TCE list (the transit search performed by the *Kepler* Pipeline) and the completeness of the KOI catalog (the vetting of the TCEs). The completeness of the *Kepler* Pipeline and its search for transits has been studied by injecting transit signals into the pixels and examining what fraction are found by the *Kepler* Pipeline (Christiansen 2017; Christiansen et al. 2015, 2013a). Burke et al. (2015) applied the appropriate detection efficiency contours (Christiansen 2015) to the 50–300 d period planet candidates in the Q1–Q16 KOI catalog (Mullally et al. 2015) in order to measure the occurrence rates of small planets. However, that study was not able to account for those transit signals correctly identified by the *Kepler* Pipeline but thrown-out by the vetting process. Along with the DR25 KOI catalog, we provide a measure of the completeness of the DR25 vetting process.

Kepler light curves contain variability that is not due to planet transits or eclipsing binaries. While the reliability of *Kepler* catalogs against astrophysical false positives is mostly understood (see e.g. Morton et al. 2016), the reliability against false alarms (a term used in this paper to indicate TCEs caused by intrinsic stellar variability, over-contact binaries, or instrumental noise, i.e., anything that does not look transit-like) has not previously been measured. Instrumental noise, statistical fluctuations, poor detrending, and/or stellar variability can conspire to produce a signal that looks similar to a planet transit. When examining the smallest exoplanets in the longest orbital periods, Burke et al. (2015) demonstrated the importance of understanding the reliability of the catalog, showing that the occurrence of small, earth-like-period planets around G dwarf stars changed by a factor of ≈10 depending on the reliability of a few planet candidates. In this catalog we measure the reliability of the reported planet candidates against this instrumental and stellar noise.

The completeness of the vetting process is measured by vetting thousands of injected transits found by the *Kepler* Pipeline. Catalog reliability is measured by vetting signals found in scrambled and inverted *Kepler* light curves and counting the fraction of simulated false alarms that are dispositioned as planet candidates. This desire to vet both the real and simulated TCEs in a reproducible and consistent manner demands an entirely automated method for vetting the TCEs.

Automated vetting was introduced in the Q1–Q16 KOI catalog (Mullally et al. 2015) with the Centroid Robovetter and was then extended to all aspects of the vetting process for the DR24 KOI catalog (Coughlin et al. 2016). Because of this automation, the DR24 catalog was the first with a measure of completeness that extended to all parts of the search, from pixels to planet candidates. Now, with the DR25 KOI catalog and simulated false alarms, we also provide a measure of how effective the vetting techniques are at identifying noise signals and translate that into a measure of the catalog reliability. As a result, the DR25 KOI catalog is the first to explicitly balance the gains in completeness against the loss of reliability, instead of always erring on the side of high completeness.

1.2. *Terms and Acronyms*

We try to avoid unnecessary acronyms and abbreviations, but a few are required to efficiently discuss this catalog. Here we itemize those terms and abbreviations that are specific to this paper and are used repeatedly. The list is short enough that we choose to group them by meaning instead of alphabetically.

TCE: Threshold Crossing Event. Periodic signals identified by the transiting planet search (TPS) module of the *Kepler* Pipeline (Jenkins 2017b).

obsTCE: Observed TCEs. TCEs found by searching the observed DR25 *Kepler* data and reported in Twicken et al. (2016).

injTCE: Injected TCEs. TCEs found that match a known, injected transit signal (Christiansen 2017).

invTCE: Inverted TCEs. TCEs found when searching the inverted data set in order to simulate instrumental false alarms (Coughlin 2017b).

scrTCE: Scrambled TCEs. TCEs found when searching the scrambled data set in order to simulate instrumental false alarms (Coughlin 2017b).

TPS: Transiting Planet Search module. This module of the *Kepler* Pipeline performs the search for planet candidates. Significant, periodic events are identified by TPS and turned into TCEs.

DV: Data Validation. Named after the module of the *Kepler* Pipeline (Jenkins 2017b) which characterizes the transits and outputs one of the detrended light curves used by the Robovetter metrics. DV also created two sets of transit fits: original and supplemental (§2.4).

ALT: Alternative. As an alternative to the DV detrending, the *Kepler* Pipeline implements a detrending method that uses the methods of Garcia (2010) and the out-of-transit points in the presearch data conditioned (PDC) light curves to detrend the data. The *Kepler* Pipeline performs a trapezoidal fit to the folded transit on the ALT detrended light curves.

MES: Multiple Event Statistic. A statistic that measures the combined significance of all of the observed transits in the detrended, whitened light curve assuming a linear ephemeris (Jenkins 2002).

KOI: Kepler Object of Interest. Periodic, transit-like events that are significant enough to warrant further review. A KOI is identified with a KOI number and can be dispositioned as a planet candidate or a false positive. The DR25 KOIs are a subset of the DR25 obsTCEs.

PC: Planet Candidate. A TCE or KOI that passes all of the Robovetter false positive identification tests. Planet candidates should not be confused with confirmed planets where further analysis has shown that the transiting planet model is overwhelmingly the most likely astrophysical cause for the periodic dips in the *Kepler* light curve.

FP: False Positive. A TCE or KOI that fails one or more of the Robovetter tests. Notice that the term includes all types of signals found in the TCE lists that are not caused by a transiting exoplanet, including eclipsing binaries and false alarms.

MCMC: Markov chain Monte Carlo. This refers to transit fits which employ a MCMC algorithm in order to provide robust errors for fitted model parameters for all KOIs (Hoffman & Rowe 2017).

1.3. *Summary and Outline of the Paper*

The DR25 KOI catalog is a uniformly-vetted list of planet candidates and false positives found by searching the DR25 *Kepler* light curves and includes a measure of the catalog completeness and reliability. In the brief outline that follows we highlight how the catalog was assembled, how we measure the completeness and reliability, and discuss those aspects of the process that are different from the DR24 KOI catalog (Coughlin et al. 2016).

In §2.1 we describe the observed TCEs (obsTCEs) which are the periodic signals found in the actual *Kepler* light curves. For reference, we also compare them to the DR24 TCEs. To create the simulated data sets necessary to measure the vetting completeness and the catalog reliability, we ran the *Kepler* Pipeline on light curves that either contained injected transits, were inverted, or were scrambled. This creates injTCEs, invTCEs, and scrTCEs, respectively (see §2.3).

We then created and tuned a Robovetter to vet all the different sets of TCEs. §3 describes the metrics and the logic used to disposition TCEs into PCs and FPs. Because the DR25 obsTCE population was significantly different than the DR24 obsTCEs, we developed new metrics to separate the PCs from the FPs (see Appendix A for the details on how each metric operates.) Several new metrics examine the individual transits for evidence of instrumental noise (see §A.3.7.) As in the DR24 KOI catalog, we group FPs into four categories (§4) and provide minor false positive flags (Appendix B) to indicate why the Robovetter decided to pass or fail a TCE. New to this catalog is the addition of a disposition score (§3.2) that gives users a measure of the Robovetter's confidence in each disposition.

Unlike previous catalogs, for the DR25 KOI catalog the choice of planet candidate versus false positive is no longer based on the philosophy of "innocent until proven guilty". We accept certain amounts of collateral damage (i.e., exoplanets dispositioned as FP) in order to achieve a catalog that is uniformly vetted and has acceptable levels of both completeness and reliability, especially for the long period and low signal-to-noise PCs. In §5 we discuss how we tuned the Robovetter using the simulated TCEs as populations of true planet candidates and true false alarms. We provide the Robovetter source code and all the Robovetter metrics for all of the sets of TCEs (obsTCEs, injTCEs, invTCEs, and scrTCEs) to enable users to create a catalog tuned for other regions of parameter space if their scientific goals require it.

We assemble the catalog (§6) by federating to previously known KOIs before creating new KOIs. Then to provide planet parameters, each KOI is fit with a transit model which uses a Markov Chain Monte Carlo (MCMC) algorithm to provide error estimates for each fitted parameter (§6.3). In §7 we summarize the catalog and discuss the performance of the vetting using the injTCE, invTCE, and scrTCE sets. We show that both decrease significantly with decreasing number of transits and decreasing signal-to-noise. We then discuss how one may use the disposition scores to identify the highest quality candidates, especially at long periods (§7.3.4.) We conclude that not all declared planet candidates in our catalog are actually astrophysical transits, but we can measure what fraction are caused by stellar and instrumental noise. Because of the interest in terrestrial, temperate planets, we examine the high quality, small candidates in the habitable zone in §7.5. Finally, in §8 we give an overview of what must be considered when using this catalog to measure accurate exoplanet occur-

rence rates, including what information is available in other *Kepler* products to do this work.

2. THE Q1–Q17 DR25 TCES

2.1. *Observed TCEs*

As with the previous three Kepler KOI catalogs (Coughlin et al. 2016; Mullally et al. 2015; Rowe et al. 2015a), the population of events that were used to create KOIs and planet candidates are known as obsTCEs. These are periodic reductions of flux in the light curve that were found by the TPS module and evaluated by the DV module of the *Kepler* Pipeline (Jenkins 2017b)[4]. The DR25 obsTCEs were created by running the SOC 9.3 version of the *Kepler* Pipeline on the DR25, Q1–Q17 *Kepler* time-series. For a thorough discussion of the DR25 TCEs and on the pipeline's search see Twicken et al. (2016).

The DR25 obsTCEs, their ephemerides, and the metrics calculated by the pipeline are available at the NASA Exoplanet Archive (Akeson et al. 2013). In this paper we endeavor to disposition these signals into planet candidates and false positives. Because the obsTCEs act as the input to our catalog, we first describe some of their properties as a whole and reflect on how they are different from the obsTCE populations found with previous searches.

We have plotted the distribution of the 32,534 obsTCEs in terms of period in Figure 1. Notice that there is an excessive number of short and long period obsTCEs compared to the number of expected transiting planets. Not shown, but worth noting is that the number of obsTCEs increases with decreasing MES.

As with previous catalogs, the short period (< 10 d) excess is dominated by true variability of stars due to both intrinsic stellar variability (e.g., spots or pulsations) and contact/near-contact eclipsing binaries. The long period excess is dominated by instrumental noise. For example, a decrease in flux following a cosmic ray hit (known as an SPSD; Van Cleve et al. 2016a), can match up with other decrements in flux to produce a TCE. Also, image artifacts known as rolling-bands are very strong on some channels (see §6.7 of Van Cleve & Caldwell 2016) and since the spacecraft rolls approximately every 90 d, causing a star to move on/off a *Kepler* detector with significant rolling band noise, these variations can easily line up to produce TCEs at *Kepler*'s heliocentric orbital period (≈372 days, 2.57 in log-space). This is the reason for the largest spike in the obsTCE population seen in Figure 1. The narrow spike at 459 days (2.66 in log-space) in the DR24 obsTCE distribution is caused by edge-effects near three equally spaced data gaps in the DR24 data processing. The short period spikes in the distribution of both the DR25 and DR24 obsTCEs is caused by contamination by bright variable stars (see §A.6 and Coughlin et al. 2014).

Generally, the excess of long period TCEs is significantly larger than it was in the DR24 TCE catalog (Seader et al. 2015), also seen in Figure 1. Most likely, this is because DR24 implemented an aggressive veto known as the bootstrap metric (Seader et al. 2015). For DR25 this metric was calculated, but was not used as a veto. Also, other vetoes were made less strict causing more TCEs across all periods to be created.

To summarize, for DR25 the number of false signals among the obsTCEs is dramatically larger than in any previous catalog. This was done on purpose in order to increase the Pipeline completeness by allow more transiting exoplanets to be made into obsTCEs.

2.2. *Rogue TCEs*

The DR25 TCE table at the NASA Exoplanet Archive contains 32,534 obsTCEs and 1498 rogue TCEs[5] for a total of 34,032. The rogue TCEs were created because of a bug in the Kepler pipeline which failed to veto certain TCEs with three transit events. This bug was not in place when characterizing the Pipeline using flux-level transit injection (see Burke & Catanzarite 2017b,a) and because the primary purpose of this catalog is to be able to accurately calculate occurrence rates, we do not use the rogue TCEs in the creation and analysis of the DR25 KOI catalog. Also note that all of the TCE populations (observed, injection, inversion, and scrambling, see the next section) had rogue TCEs that were removed prior to analysis. The creation and analysis of this KOI catalog only rely on the non-rogue TCEs. Although they are not analyzed in this study we encourage the community to examine the designated rogue TCEs as the list does contain some of the longest period events detected by *Kepler*.

2.3. *Simulated TCEs*

In order to measure the performance of the Robovetter and the *Kepler* Pipeline, we created simulated transits, simulated false positives, and simulated false alarms. The simulated transits are created by injecting transit signals into the pixels of the original data. The simulated false positives were created by injecting eclipsing binary signals and positionally off-target transit signals into the pixels of the original data (see Coughlin 2017b and Christiansen 2017 for more information). The simulated false alarms were created in two separate ways: by inverting the light curves, and by scrambling the sequence of cadences in the time series. The TCEs that resulted from these simulated data are available at the Exoplanet Archive on the Kepler simulated data page.[6]

[4] The source code of the entire Pipeline is available at https://github.com/nasa/kepler-pipeline

[5] See the tce_rogue_flag column in the DR25 TCE table at the exoplanet archive.

[6] https://exoplanetarchive.ipac.caltech.edu/docs/KeplerSimulated.html

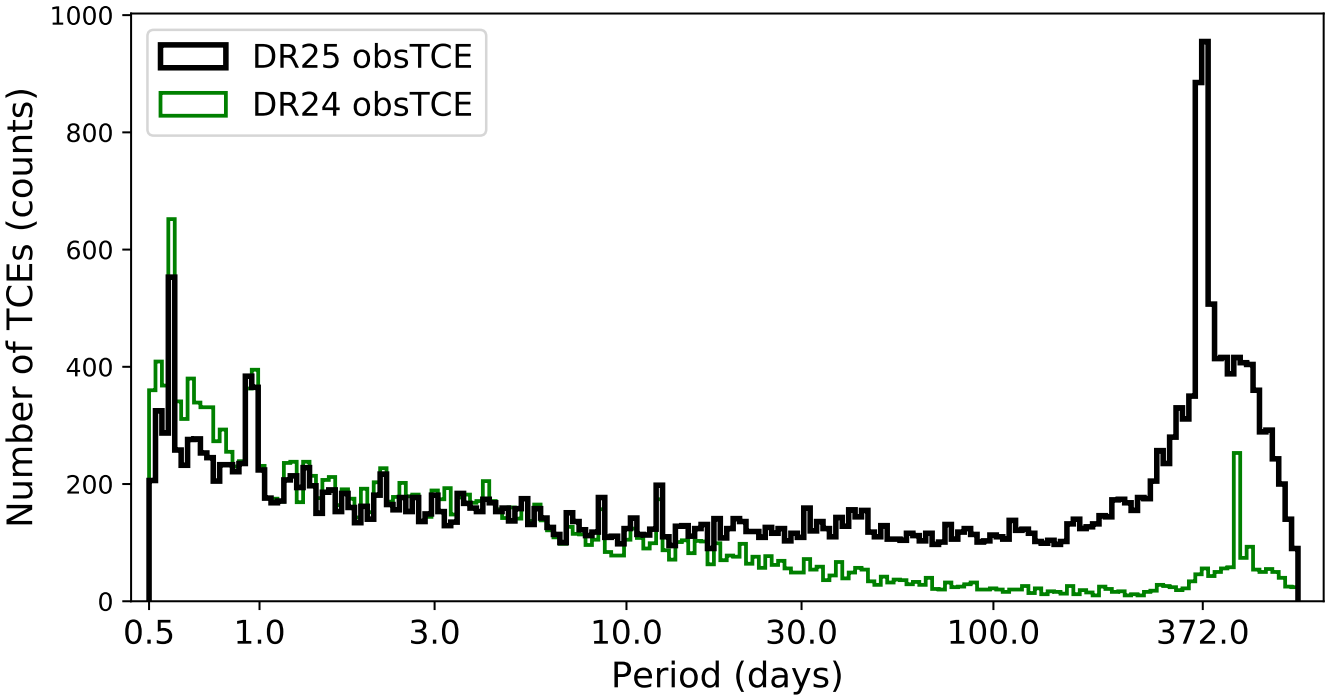

Figure 1. Histogram of the period in days of the DR25 obsTCEs (black) using uniform bin space in the base ten logarithm of the period. The DR24 catalog obsTCEs (Seader et al. 2015) are shown in green for comparison. The number of long-period TCEs is much larger for DR25 and includes a large spike in the number of TCEs at the orbital period of the spacecraft (372 days). The long and short period spikes for both distributions are discussed in §2.1.

2.3.1. *True Transits – Injection*

We empirically measure the completeness of the *Kepler* Pipeline and the subsequent vetting by injecting a suite of simulated transiting planet signals into the calibrated pixel data and observing their recovery, as was done for previous versions of the *Kepler* Pipeline (Christiansen et al. 2013a; Christiansen 2015; Christiansen et al. 2016). The full analysis of the DR25 injections are described in detail in Christiansen (2017). In order to understand the completeness of the Robovetter, we use the on-target injections (Group 1 in Christiansen 2017); we briefly describe their properties here. For each of the 146,294 targets, we generate a model transit signal using the Mandel & Agol (2002) formulation, with parameters drawn from the following uniform distributions: orbital periods from 0.5–500 days (0.5–100 days for M dwarf targets), planet radii from 0.25–7 $R_\oplus$ (0.25–4 $R_\oplus$ for M dwarf targets), and impact parameters from 0–1. After some re-distribution in planet radius to ensure sufficient coverage where the *Kepler* Pipeline is fully incomplete (0% recovery) to fully complete (100% recovery), 50% of the injections have planet radii below 2 $R_\oplus$ and 90% below 40 $R_\oplus$. The signals are injected into the calibrated pixels, and then processed through the remaining components of the *Kepler* Pipeline in an identical fashion to the original data. Any detected signals are subjected to the same scrutiny by the Pipeline and the Robovetter as the original data. By measuring the fraction of injections that were successfully recovered by the Pipeline and called a PC by the Robovetter with any given set of parameters (e.g., orbital period and planet radius), we can then correct the number of candidates found with those parameters to the number that are truly present in the data. While the observed population of true transiting planets is heavily concentrated towards short periods, we chose the 0.5–500 day uniform period distribution of injections because more long-period, low signal-to-noise transits are both not recovered and not vetted correctly — injecting more of these hard-to-find, long-period planets ensures that we can measure the Pipeline and Robovetter completeness. In this paper we use the set of on-target, injected planets that were recovered by the *Kepler* Pipeline (the injTCEs, whose period distribution is shown in Figure 2) to measure the performance of the Robovetter. Accurate measurement of the Robovetter performance is limited to those types of transits injected and recovered.

It is worth noting that the injections do not completely emulate all astrophysical variations produced by a planet transiting a star. For instance, the injected model includes limb-darkening, but not the occultation of stellar pulsations or granulation, which has been shown to cause a small, but non-negligible, error source

on measured transit depth (Chiavassa et al. 2017) for high signal-to-noise transits.

2.3.2. *False Alarms – Inverted and Scrambled*

To create realistic false alarms that have noise properties similar to our obsTCEs, we inverted the light curves (i.e., multiplied the normalized, zero-mean flux values by negative one) before searching for transit signals. Because the pipeline is only looking for transit-like (negative) dips in the light curve, the true exoplanet transits should no longer be found. However, quasi-sinusoidal signals due to instrumental noise, contact and near-contact binaries, and stellar variability can still create detections. In order for inversion to exactly reproduce the false alarm population, the false alarms would need to be perfectly symmetric (in shape and frequency) under flux inversion, which is not true. For example, stellar oscillations and star-spots are not sine waves and SPSDs will not appear the same under inversion. However, the rolling band noise that is significant on many of *Kepler*'s channels is mostly symmetric. The period distribution of these invTCEs is shown in Figure 2. The distribution qualitatively emulates those seen in the obsTCEs; however there are only ∼60% as many. This is because the population does not include the exoplanets nor the eclipsing binaries, but it is also because many of the sources of false alarms are not symmetric under inversion. The one-year spike is clearly seen, but is not as large as we might expect, likely because the broad long-period hump present in the DR25 obsTCE distribution is mostly missing from the invTCE distribution. We explore the similarity of the invTCEs to obsTCEs in more detail in §4.2.

Another method to create false alarms is to scramble the order of the data. The requirement is to scramble the data enough to lose the coherency of the binary stars and exoplanet transits, but to keep the coherency of the instrumental and stellar noise that plagues the *Kepler* data set. Our approach was to scramble the data in coherent chunks of one year. The fourth year of data (Q13–Q16) was moved to the start of the light curve, followed by the third year (Q9–Q12), then the second (Q5–Q8), and finally the first (Q1–Q4). Q17 remained at the end. Within each year, the order of the data did not change. Notice that in this configuration each quarter remains in the correct *Kepler* season preserving the yearly artifacts produced by the spacecraft.

Two additional scrambling runs of the data, with different scrambling orders than described above, were performed and run through the *Kepler* pipeline and Robovetter, but are not discussed in this paper, as they were produced after the analysis for this paper was complete. These runs could be very useful in improving the reliability measurements of the DR25 catalog — see Coughlin 2017b for more information.

2.3.3. *Cleaning Inversion and Scrambling*

As will be described in §4.1, we want to use the invTCE and scrTCE sets to measure the reliability of the DR25 catalog against instrumental and stellar noise. In order to do that well, we need to remove signals found in these sets that are not typical of those in our obsTCE set. For inversion, there are astrophysical events that look similar to an inverted eclipse, for example the self-lensing binary star, KOI 3278.01 (Kruse & Agol 2014), and Heartbeat binaries (Thompson et al. 2012). With the assistance of published systems and early runs of the Robovetter, we identified any invTCE that could be one of these types of astrophysical events; 54 systems were identified in total. Also, the shoulders of inverted eclipsing binary stars and high signal-to-noise KOIs are found by the Pipeline, but are not the type of false alarm we were trying to reproduce, since they have no corresponding false alarm in the original, un-inverted light curves. We remove any invTCEs that were found on stars that had 1) one of the identified astrophysical events, 2) a detached eclipsing binary listed in Kirk et al. (2016) with morphology values larger than 0.6, or 3) a known KOI. After cleaning, we are left with 14953 invTCEs; their distribution is plotted in the top of Figure 2.

For the scrambled data, we do not have to worry about the astrophysical events that emulate inverted transits, but we do have to worry about triggering on true transits that have been rearranged to line up with noise. For this reason we remove from the scrTCE population all that were found on a star with a known eclipsing binary (Kirk et al. 2016), or on an identified KOI. The result is 13782 scrTCEs; their distribution is plotted in the middle panel of Figure 2. This will not remove all possible sources of astrophysical transits. Systems with only two transits (which would not be made into KOIs), or systems with single transits from several orbiting bodies would not be identified in this way. For example, KIC 3542116 was identified by Rappaport et al. (2017) as a star with possible exocomets, and it is a scrTCE dispositioned as an FP. We expect the effect of not removing these unusual events to be negligible on our reliability measurements relative to other systematic differences between the obsTCEs and the scrTCEs.

After cleaning the invTCEs and scrTCEs, the number of scrTCEs at periods longer than 200 d closely matches the size and shape of the obsTCE distribution, except for the one-year spike. The one-year spike is well represented by the invTCEs. The distribution of the combined invTCE and scrTCE data sets, as shown in the middle plot of Figure 2, qualitatively matches the relative frequency of false alarms present in the DR25 obsTCE population. Tables 1 and 2 lists those invTCEs and scrTCEs that we used when calculating the false alarm effectiveness and false alarm reliability of the PCs.

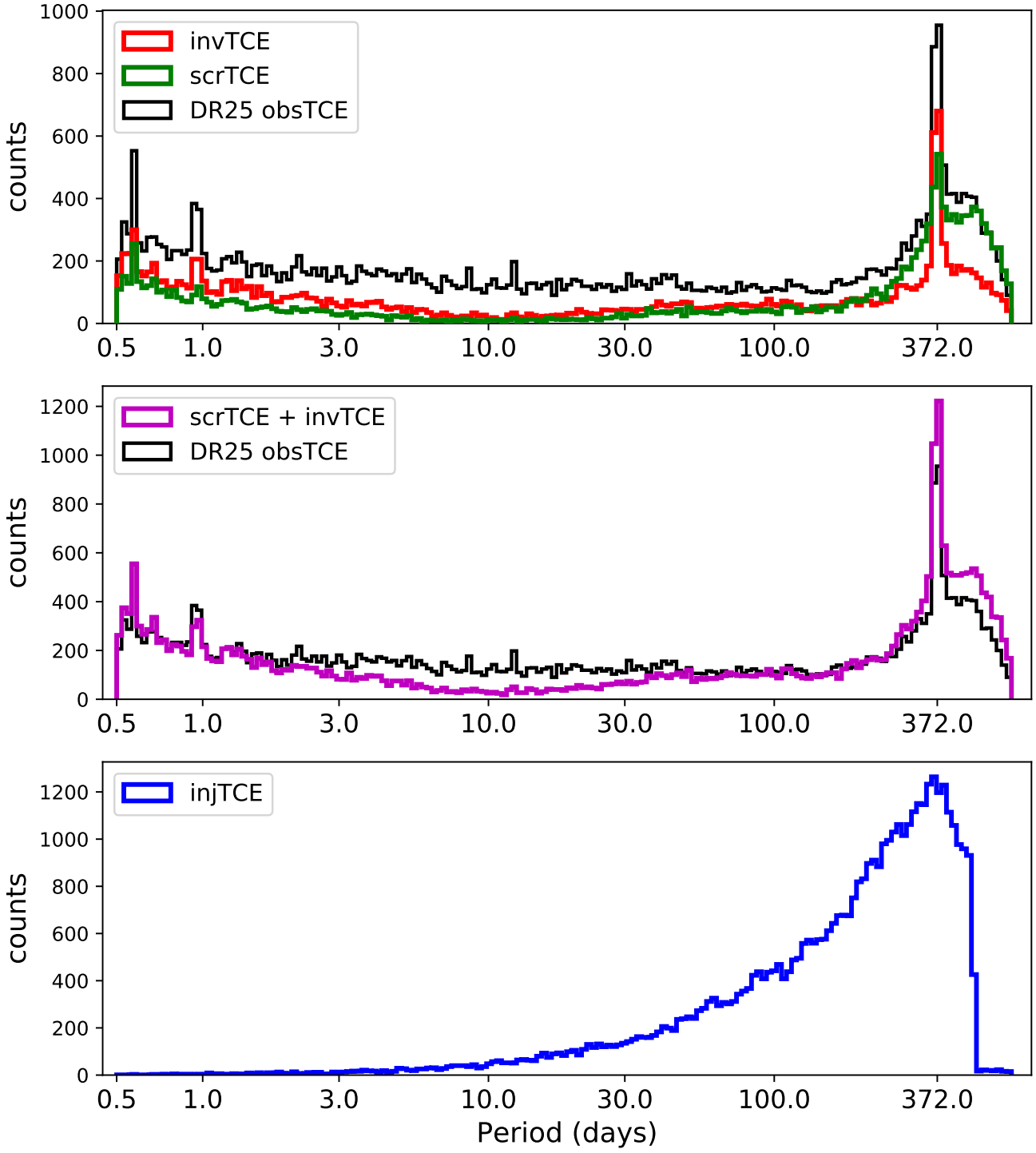

Figure 2. Histogram of the period in days of the cleaned invTCEs (top, red), the cleaned scrTCEs (top, green), and injTCEs (bottom, blue) in uniform, base-ten logarithmic spacing. The middle plot shows the union of the invTCEs and the scrTCEs in magenta. The DR25 obsTCEs are shown for comparison on the top two figures in black. At shorter periods (< 30 days) in the top figure, the difference between the simulated false alarm sets and the observed data represents the number of transit-like KOIs; at longer periods we primarily expect false alarms. Notice that the invTCEs do a better job of reproducing the one-year spike, but the scrTCEs better reproduce the long-period hump. Because the injTCEs are dominated by long-period events (significantly more long-period events were injected), we are better able to measure the Robovetter completeness for long-period planets than short-period planets.

Table 1. invTCEs used in the analysis of catalog reliability

TCE-ID (KIC-PN)	Period days	MES	Disposition PC/FP
000892667-01	2.261809	7.911006	FP
000892667-02	155.733356	10.087069	FP
000892667-03	114.542735	9.612742	FP
000892667-04	144.397127	8.998353	FP
000892667-05	84.142047	7.590044	FP
000893209-01	424.745158	9.106225	FP
001026133-01	1.346275	10.224972	FP
001026294-01	0.779676	8.503883	FP
001160891-01	0.940485	12.176910	FP
001160891-02	0.940446	13.552523	FP
001162150-01	1.130533	11.090898	FP
001162150-02	0.833482	8.282225	FP
001162150-03	8.114960	11.956621	FP
001162150-04	7.074370	14.518677	FP
001162150-05	5.966962	16.252800	FP
...	...	...	...

Note—The first column is the TCE-ID and is formed using the KIC Identification number and the TCE planet number (PN). This table is published in its entirety in the machine-readable format. A portion is shown here for guidance regarding its form and content.

Table 2. scrTCEs used in the analysis of catalog reliability

TCE-ID (KIC-PN)	Period days	MES	Disposition PC/FP
000757099-01	0.725365	8.832907	FP
000892376-01	317.579997	11.805184	FP
000892376-02	1.532301	11.532692	FP
000892376-03	193.684366	14.835271	FP
000892376-04	432.870540	11.373951	FP
000892376-05	267.093312	10.308785	FP
000892376-06	1.531632	10.454597	FP
000893004-01	399.722285	7.240176	FP
000893507-02	504.629640	15.434824	FP
000893507-03	308.546946	12.190248	FP
000893507-04	549.804329	12.712417	FP
000893507-05	207.349237	11.017911	FP
000893647-01	527.190559	13.424537	FP
000893647-02	558.164884	13.531707	FP
000893647-03	360.260977	9.600089	FP
...	...	...	...

Note—The first column is the TCE-ID and is formed using the KIC Identification number and the TCE planet number (PN). This table is published in its entirety in the machine-readable format. A portion is shown here for guidance regarding its form and content.

2.4. *TCE Transit Fits*

The creation of this KOI catalog depends on four different transit fits: 1) the original DV transit fits, 2) the trapezoidal fits performed on the ALT Garcia (2010) detrended light curves, 3) the supplemental DV transit fits, and 4) the MCMC fits (see §6.3). The *Kepler* Pipeline fits each TCE with a Mandel & Agol (2002) transit model using Claret (2000) limb darkening parameters. After the transit searches were performed for the observed, injected, scrambled, and inverted TCEs, we discovered that the transit fit portion of DV was seeded with a high impact parameter model that caused the final fits to be biased towards large values, causing the planet radii to be systematically too large (for further information see Christiansen 2017 and Coughlin 2017b). Since a consistent set of reliable transit fits are required for all TCEs, we refit the transits. The same DV transit fitting code was corrected for the bug and seeded with the *Kepler* identification number, period, epoch, and MES of the original detection. These "supplemental" DV fits do not have the same impact parameter bias as the original. Sometimes the DV fitter fails to converge and in these cases we were not able to obtain a supplemental DV transit fit, causing us to fall back on the original DV fit. Also, at times the epoch wanders too far from the original detection; in these cases we do not consider it to be a successful fit and again fall back on the original fit.

Because the bug in the transit fits was only discovered after all of the metrics for the Robovetter were run, the original DV and trapezoidal fits were used to disposition all of the sets of TCEs. These are the same fits that are available for the obsTCEs in the DR25 TCE table at the NASA Exoplanet Archive. Most Robovetter metrics are agnostic to the parameters of the fit, and so the supplemental DV fits would only change a few of the Robovetter decisions. While the Robovetter itself runs in a few minutes, several of the metrics used by the Robovetter (see Appendix A) require weeks to compute, so we chose not to update the metrics in order to achieve this minimal improvement. And for all sets of TCEs, the original DV fits are listed in the Robovetter input files[7]. The supplemental fits are used to understand the completeness and reliability of the catalog as a function of fitted parameters (such as planet radii or insolation

[7] Robovetter input files have the format kplr_dr25_obs_robovetter_input.tar.gz and can be found in the Robovetter github repository, https://github.com/nasa/kepler-robovetter

flux). For all sets of TCEs, the supplemental DV fits are available as part of the Robovetter results tables linked from the TCE documentation page[8] for the obsTCEs and from the simulated data page[9] (see Christiansen 2017; Coughlin 2017b) for the injected, inverted, and scrambled TCEs. The MCMC fits are only provided for the KOI population and are available in the DR25 KOI table[10] at the NASA Exoplanet Archive. The MCMC fits have no consistent offset from the supplemental DV fits. To show this, we plot the planet radii derived from the two types of fits for the planet candidates in DR25 and show the distribution of fractional change in planet radii; see Figure 3. The median value of the fractional change is 0.7% with a standard deviation of 18%. While individual systems disagree, there is no offset in planet radii between the two populations. The supplemental DV fitted radii and MCMC fitted radii agree within 1-sigma of the combined error bar (i.e., the square-root of the sum of the squared errors) for 78% of the KOIs and 93.4% of PCs (only 1.8% of PC's radii differ by more than 3-sigma). The differences are caused by discrepancies in the detrending and because the MCMC fits include a non-linear ephemeris in its model when appropriate (i.e., to account for transit-timing variations).

2.5. *Stellar Catalog*

Some of the derived parameters from transit fits (e.g., planetary radius and insolation flux) of the TCEs and KOIs rely critically on the accuracy of the stellar properties (e.g., radii, mass, and temperature). For all transit fits used to create this catalog we use the DR25 Q1–Q17 stellar table provided by Mathur et al. (2017), which is based on conditioning published atmospheric parameters on a grid of Dartmouth isochrones (Dotter et al. 2008). The best-available observational data for each star is used to determine the stellar parameters; e.g., asteroseismic or high-resolution spectroscopic data, when available, is used instead of broad-band photometric measurements. Typical uncertainties in this stellar catalog are ≈27% in radius, ≈17% in mass, and ≈51% in density, which is somewhat smaller than previous catalogs.

After completion of the DR25 catalog an error was discovered: the metallicities of 780 KOIs were assigned a fixed erroneous value ([Fe/H] = 0.15 dex). These targets can be identified by selecting those that have the metallicity provenance column set to "SPE90". Since radii are fairly insensitive to metallicity and the average metallicity of *Kepler* stars is close to solar, the impact of this error on stellar radii is typically less than 10% and does not significantly change the conclusions in this paper. Corrected stellar properties for these stars will be provided in an upcoming erratum to Mathur et al. (2017). The KOI catalog vetting and fits rely exclusively on the original DR25 stellar catalog information.

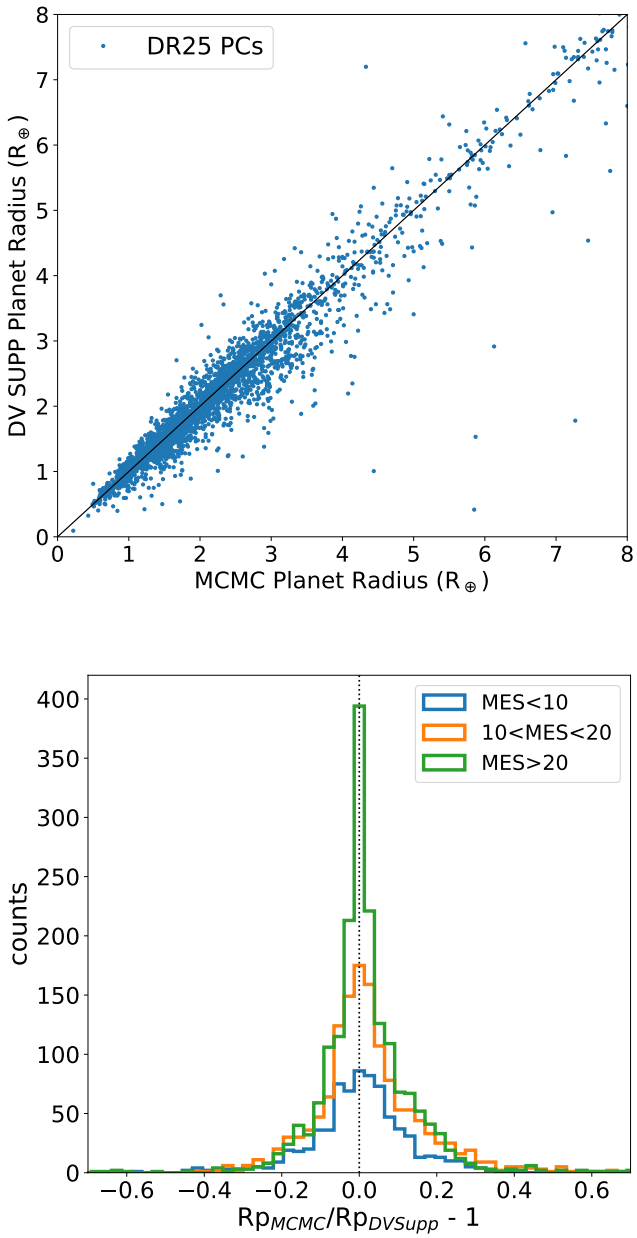

Figure 3. Top: Comparison of the DR25 PCs fitted planet radii measured by the MCMC fits and the DV supplemental fits. The 1:1 line is drawn in black. Bottom: Histogram of the difference between the MCMC fits and the DV fits for the planet candidates in different MES bins. While individual objects have different fitted values, as a group the planet radii from the two fits agree.

[8] The Robovetter results files are linked under the Q1-Q17 DR25 Information on the page https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_TCE_docs.html

[9] https://exoplanetarchive.ipac.caltech.edu/docs/KeplerSimulated.html

[10] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr25_koi

Because the stellar parameters will continue to be updated (with data from missions such as *Gaia*, Gaia Collaboration et al. 2016b,a) we perform our vetting and analysis independent of stellar properties where possible and provide the fitted information relative to the stellar properties in the KOI table. A notable exception is the limb darkening values; precise transit models require limb darkening coefficients that depends on the stellar temperature and gravity. However, limb-darkening coefficients are fairly insensitive to the most uncertain stellar parameters in the stellar properties catalog (e.g., surface gravity; Claret 2000).

3. THE ROBOVETTER: VETTING METHODS AND METRICS

The dispositioning of the TCEs as PC and FP is entirely automated and is performed by the Robovetter[11]. This code uses a variety of metrics to evaluate and disposition the TCEs.

Because the TCE population changed significantly between DR24 and DR25 (see Figure 1), the Robovetter had to be improved in order to obtain acceptable performance. Also, because we now have simulated false alarms (invTCEs and scrTCEs) and true transits (injTCEs), the Robovetter could be tuned to keep the most injTCEs and remove the most invTCEs and scrTCEs. This is a significant change from previous KOI catalogs that prioritized completeness above all else. In order to sufficiently remove the long period excess of false alarms, this Robovetter introduces new metrics that evaluate individual transits (in addition to the phase-folded transits), expanding the work that the code Marshall (Mullally et al. 2016) performed for the DR24 KOI catalog.

Because most of the Robovetter tests and metrics changed between DR24 and DR25, we fully describe all of the metrics. In this section we summarize the important aspects of the Robovetter logic and only provide a list of each test's purpose. The details of these metrics, and more details on the Robovetter logic, can be found in Appendix A. We close this section by explaining the creation of the "disposition score", a number which conveys the confidence in the Robovetter's disposition.

3.1. *Summary of the Robovetter*

In Figure 4 we present a flowchart that outlines our robotic vetting procedure. Each TCE is subjected to a series of "yes" or "no" questions (represented by diamonds) that either disposition it into one or more of the four FP categories, or else disposition it as a PC. Behind each question is a series of more specific questions, each answered by quantitative tests.

First, if the TCE under investigation is not the first in the system, the Robovetter checks if the TCE corresponds to a secondary eclipse associated with an already examined TCE in that system. If not, the Robovetter then checks if the TCE is transit-like. If it is transit-like, the Robovetter then looks for the presence of a secondary eclipse. In parallel, the Robovetter looks for evidence of a centroid offset, as well as an ephemeris match to other TCEs and variable stars in the *Kepler* field.

Similar to previous KOI catalogs (Coughlin et al. 2016; Mullally et al. 2015; Rowe et al. 2015a), the Robovetter assigns FP TCEs to one or more of the following false positive categories:

- Not Transit-Like (NT): a TCE whose light curve is not consistent with that of a transiting planet or eclipsing binary. These TCEs are usually caused by instrumental artifacts or non-eclipsing variable stars. If the Robovetter worked perfectly, all false alarms, as we have defined them in this paper, would be marked as FPs with only this Not Transit-Like flag set.

- Stellar Eclipse (SS): a TCE that is observed to have a significant secondary event, v-shaped transit profile, or out-of-eclipse variability that indicates the transit-like event is very likely caused by an eclipsing binary. Self-luminous, hot Jupiters with a visible secondary eclipse are also in this category, but are still given a disposition of PC. In previous KOI catalogs this flag was known as Significant Secondary.

- Centroid Offset (CO): a TCE whose signal is observed to originate from a source other than the target star, based on examination of the pixel-level data.

- Ephemeris Match Indicates Contamination (EC): a TCE that has the same period and epoch as another object, and is not the true source of the signal given the relative magnitudes, locations, and signal amplitudes of the two objects. See Coughlin (2014).

The specific tests that caused the TCE to fail are specified by minor flags. These flags are described in Appendix B and are available for all FPs. Table 3 gives a summary of the specific tests run by the Robovetter when evaluating a TCE. The table lists the false positive category (NT, SS, CO or EC) of the test and also which minor flags are set by that test. Note that there are several informative minor flags, which are listed in Appendix B, but are not listed in Table 3 because they do not change the disposition of a TCE. Also, Appendix B tabulates how often each minor flag was set to help understand the frequency of each type of FP.

New to this Robovetter are several tests that look at individual transits. The tests are named after the code

[11] https://github.com/nasa/kepler-robovetter

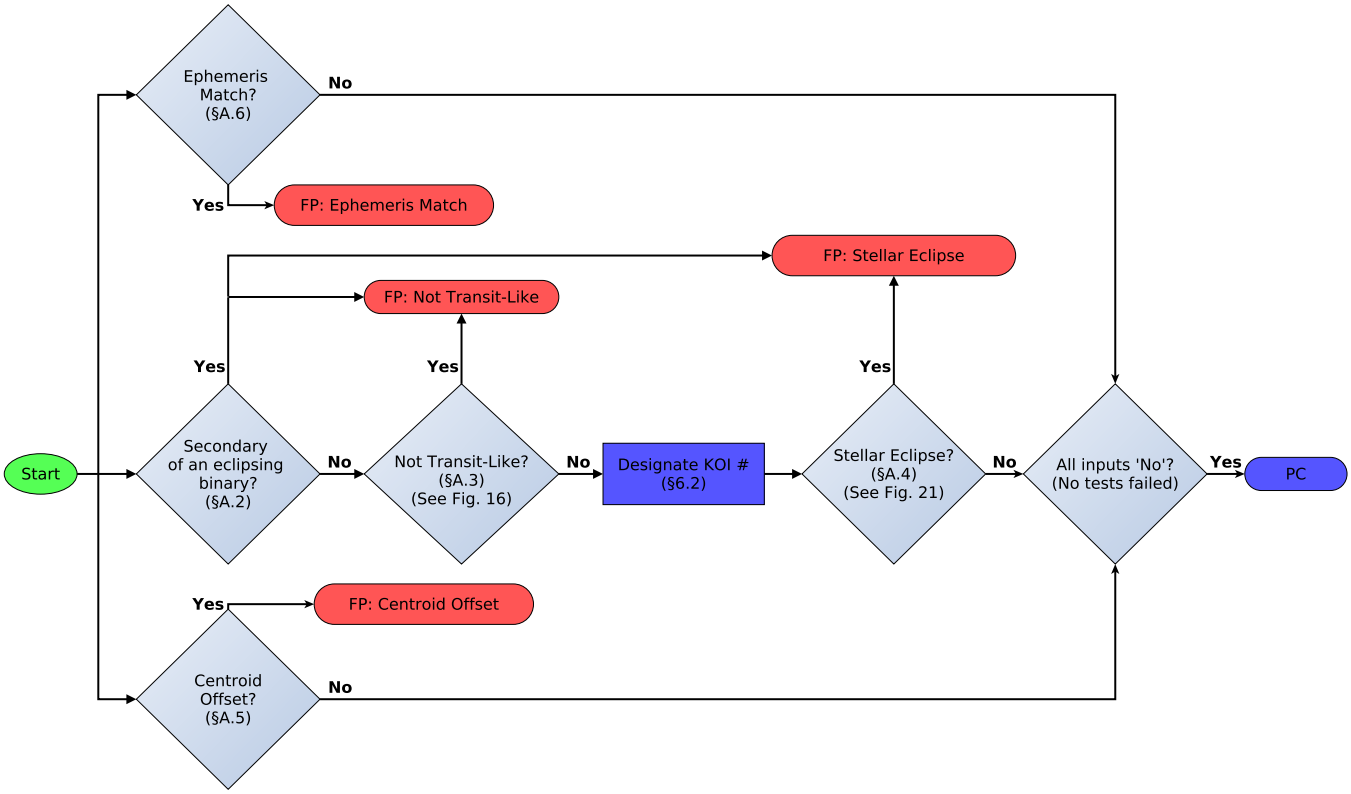

Figure 4. Overview flowchart of the Robovetter. Diamonds represent "yes" or "no" decisions that are made with quantitative metrics. A TCE is dispositioned as an FP if it fails any test (a "yes" decision) and is placed in one or more of the FP categories. (A TCE that is identified as being the secondary eclipse of a system is placed in both the Not Transit-Like and Stellar Eclipse categories.) If a TCE passes all tests (a "no" decision for all tests) it is dispositioned as a PC. The section numbers on each component correspond to the sections in this paper where these tests are discussed. More in-depth flowcharts are provided for the not transit-like and stellar eclipse modules in Figures 16 and 21.

that calculates the relevant metric and are called: Rubble, Marshall, Chases, Skye, Zuma, and Tracker. Each metric only identifies which transits can be considered "bad", or not sufficiently transit-like. The Robovetter only fails the TCE if the number of remaining good transits is less than three, or if the recalculated MES, using only the good transits, drops below 7.1.

Another noteworthy update to the Robovetter in DR25 is the introduction of the v-shape metric, originally introduced in Batalha et al. (2013). The intent is to remove likely eclipsing binaries which do not show significant secondary eclipses by looking at the shape and depth of the transit (see §A.4.3).

3.2. *Disposition Scores*

We introduce a new feature to this catalog called the Disposition Score. Essentially the disposition score is a value between 0 and 1 that indicates the confidence in a disposition provided by the Robovetter. A higher value indicates more confidence that a TCE is a PC, regardless of the disposition it was given. This feature allows one to select the highest quality PCs by ranking KOIs via the disposition score, for both use in selecting samples for occurrence rate calculations and prioritizing individual objects for follow-up. *We stress that the disposition score does not map directly to a probability that the signal is a planet.* However, in §7.3.4 we discuss how the disposition score can be used to adjust the reliability of a sample.

The disposition score was calculated by wrapping the Robovetter in a Monte Carlo routine. In each Monte Carlo iteration, for each TCE, new values are chosen for most of the Robovetter input metrics by drawing from an asymmetric Gaussian distribution [12] centered on the nominal value. The Robovetter then dispositions each TCE given the new values for its metrics. The disposition score is simply the fraction of Monte Carlo iterations that result in a disposition of PC. (We used 10,000 iterations for the results in this catalog.) For example, if a TCE that is initially dispositioned as a PC has several metrics that are just barely on the passing side of their Robovetter thresholds, in many iterations at least

[12] The asymmetric Gaussian distribution is created so that either side of the central value follows a Gaussian, each with a different standard deviation.

Table 3. Summary of the DR25 Robovetter tests

Test Name	Section	Major Flags	Minor Flags	Brief Description
Is Secondary	A.2	NT SS	IS_SEC_TCE	The TCE is a secondary eclipse.
LPP Metric	A.3.1	NT	LPP_DV LPP_ALT	The TCE is not transit-shaped.
SWEET NTL	A.3.2	NT	SWEET_NTL	The TCE is sinusoidal.
TCE Chases	A.3.3	NT	ALL_TRANS_CHASES	The individual TCE events are not uniquely significant.
MS_1	A.3.4	NT	MOD_NONUNIQ_DV MOD_NONUNIQ_ALT	The TCE is not significant compared to red noise.
MS_2	A.3.4	NT	MOD_TER_DV MOD_TER_ALT	Negative event in phased flux as significant as TCE.
MS_3	A.3.4	NT	MOD_POS_DV MOD_POS_ALT	Positive event in phased flux as significant as TCE.
Max SES to MES	A.3.5	NT	INCONSISTENT_TRANS	The TCE is dominated by a single transit event.
Same Period	A.3.6	NT	SAME_NTL_PERIOD	Has same period as a previous not transit-like TCE.
Individual Transits	A.3.7	NT	INDIV_TRANS_	Has < 3 good transits and recalculated MES < 7.1.
Rubble	A.3.7.1	···	INDIV_TRANS_RUBBLE	Transit does not contain enough cadences.
Marshall	A.3.7.2	···	INDIV_TRANS_MARSHALL	Transit shape more closely matches a known artifact.
Chases	A.3.7.3	···	INDIV_TRANS_CHASES	Transit event is not significant.
Skye	A.3.7.4	···	INDIV_TRANS_SKYE	Transit time is correlated with other TCE transits.
Zuma	A.3.7.5	···	INDIV_TRANS_ZUMA	Transit is consistent with an increase in flux.
Tracker	A.3.7.6	···	INDIV_TRANS_TRACKER	No match between fitted and discovery transit time.
Gapped Transits	A.3.8	NT	TRANS_GAPPED	The fraction of transits identified as bad is large.
MS Secondary	A.4.1.2	SS	MOD_SEC_DV MOD_SEC_ALT	A significant secondary event is detected.
Secondary TCE	A.4.1.1	SS	HAS_SEC_TCE	A subsequent TCE on this star is the secondary.
Odd Even	A.4.1.4	SS	DEPTH_ODDEVEN_DV DEPTH_ODDEVEN_ALT MOD_ODDEVEN_DV MOD_ODDEVEN_ALT	The depths of odd and even transits are different.
SWEET EB	A.4.2	SS	SWEET_EB	Out-of-phase tidal deformation is detected.
V Shape Metric	A.4.3	SS	DEEP_V_SHAPE	The transit is deep and v-shaped.
Planet Occultation[PC]	A.4.1.3	SS	PLANET_OCCULT_DV PLANET_OCCULT_ALT	Significant secondary could be planet occultation.
Planet Half Period[PC]	A.4.1.3	···	PLANET_PERIOD_IS_HALF_DV PLANET_PERIOD_IS_HALF_ALT	Planet scenario possible at half the DV period.
Resolved Offset	A.5.1	CO	CENT_RESOLVED_OFFSET	The transit occurs on a spatially resolved target.
Unresolved Offset	A.5.1	CO	CENT_UNRESOLVED_OFFSET	A shift in the centroid position occurs during transit.
Ghost Diagnostic	A.5.2	CO	HALO_GHOST	The transit strength in the halo pixels is too large.
Ephemeris Match	A.6	EC	EPHEM_MATCH	The ephemeris matches that of another source.

NOTE—More details about all of these tests and how they are used by the Robovetter can be found in the sections listed in the second column.

[PC] These tests override previous tests and will cause the TCE to become a planet candidate.

one will be perturbed across the threshold. As a result, many of the iterations will produce a false positive and the TCE will be dispositioned as a PC with a low score. Similarly, if a TCE only fails due to a single metric that was barely on the failing side of a threshold, the score may be near 0.5, indicating that it was deemed a PC in half of the iterations. Since a TCE is deemed a FP even if only one metric fails, nearly all FPs have scores less than 0.5, with most very close to 0.0. PCs have a wider distribution of scores from 0.0 to 1.0 depending on how many of their metrics fall near to the various Robovetter thresholds.

To compute the asymmetric Gaussian distribution for each metric, we examined the metric distributions for the injected on-target planet population on FGK dwarf targets. In a 20 by 20 grid in linear period space (ranging from 0.5 to 500 d) and logarithmic MES space (ranging from 7.1 to 100), we calculated the median absolute deviation (MAD) for those values greater than the median value and separately for those values less than the median value. We chose to use MAD because it is robust to outliers. MES and period were chosen as they are fundamental properties of a TCE that well characterize each metric's variation. The MAD values were then multiplied by a conversion factor of 1.4826 to put the variability on the same scale as a Gaussian standard deviation (Hampel 1974; Ruppert 2010). A two-dimensional power-law was then fitted to the 20 by 20 grid of standard deviation values, separately for the greater-than-median and less-than-median directions. With this analytical approximation for a given metric, an asymmetric Gaussian distribution can be generated for each metric for any TCE given its MES and period.

An example is shown in Figure 5 for the LPP metric (Locality Preserving Projections, see §A.3.1) using the DV detrending. The top-left plot shows the LPP values of all on-target injected planets on FGK dwarf targets as a function of period, and the top-right shows them as a function of MES. The middle-left plot shows the measured positive 1σ deviation (in the same units as the LPP metric) as a function of MES and period, and the middle-right plot shows the resulting best-fit model. The bottom plots show the same thing but for the negative 1σ deviation. As can be seen, the scatter in the LPP metric has a weak period dependence, but a strong MES dependence, due to the fact it is easier to measure the overall shape of the light curve (LPP's goal) with higher MES (signal-to-noise).

Most, but not all, of the Robovetter metrics were amenable to this approach. Specifically, the list of metrics that were perturbed in the manner above to generate the score values were: LPP (both DV and ALT), all the Model-shift metrics (MS_1, MS_2, MS_3, and MS Secondary, both DV and ALT), TCE Chases, max-SES-to-MES, the two odd/even metrics (both DV and ALT), Ghost Diagnostic, and the recomputed MES using only good transits left after the individual transit metrics.

4. CALCULATING COMPLETENESS AND RELIABILITY

We use the injTCE, scrTCE, and invTCE data sets to determine the performance of the Robovetter and to measure the completeness and the reliability of the catalog. As a reminder, the reliability we are attempting to measure is only the reliability of the catalog against false alarms and does not address the astrophysical reliability (see §8). As discussed in §2.1, the long-period obsTCEs are dominated by false alarms and so this measurement is crucial to understand the reliability of some of the most interesting candidates in our catalog.

Robovetter completeness, C, is the fraction of injected transits detected by the *Kepler* Pipeline that are passed by the Robovetter as PCs. As long as the injTCEs are representative of the observed PCs, completeness tells us what fraction of the true planets are missing from the final catalog. Completeness is calculated by dividing the number of on-target injTCEs that are dispositioned as PCs ($N_{PC_{inj}}$) by the total number of injTCEs (N_{inj}).

$$C \approx \frac{N_{\mathrm{PC_{inj}}}}{N_{\mathrm{inj}}} \tag{1}$$

If the parameter space under consideration becomes too large and there are gradients in the actual completeness, differences between the injTCEand the obsTCE populations will prevent the completeness measured with Equation 1 from matching the actual Robovetter completeness. For example, there are more long-period injTCEs than short-period ones, which is not representative of the observed PCs, the true fraction of candidates correctly dispositioned by the Robovetter is not accurately represented by binning over all periods. With this caveat in mind, we use C in this paper to indicate the value we can measure, as shown in Equation 1.

The candidate catalog reliability, R, is defined as the ratio of the number of PCs which are truly exoplanets ($T_{PC_{obs}}$) to the total number of PCs ($N_{PC_{obs}}$) from the obsTCE data set.

$$R = \frac{T_{\mathrm{PC_{obs}}}}{N_{\mathrm{PC_{obs}}}} \tag{2}$$

Calculating the reliability for a portion of the candidate catalog is not straight forward because we do not know which PCs are the true transiting exoplanets and cannot directly determine $T_{PC_{obs}}$. Instead, we use the simulated false alarm data sets to understand how often false alarms sneak past the Robovetter and contaminate our final catalog.

4.1. *Reliability Derivation*

To assess the catalog reliability against false alarms, we will assume that the scrTCEs and invTCEs are similar (in frequency and type) to the obsTCEs. One way

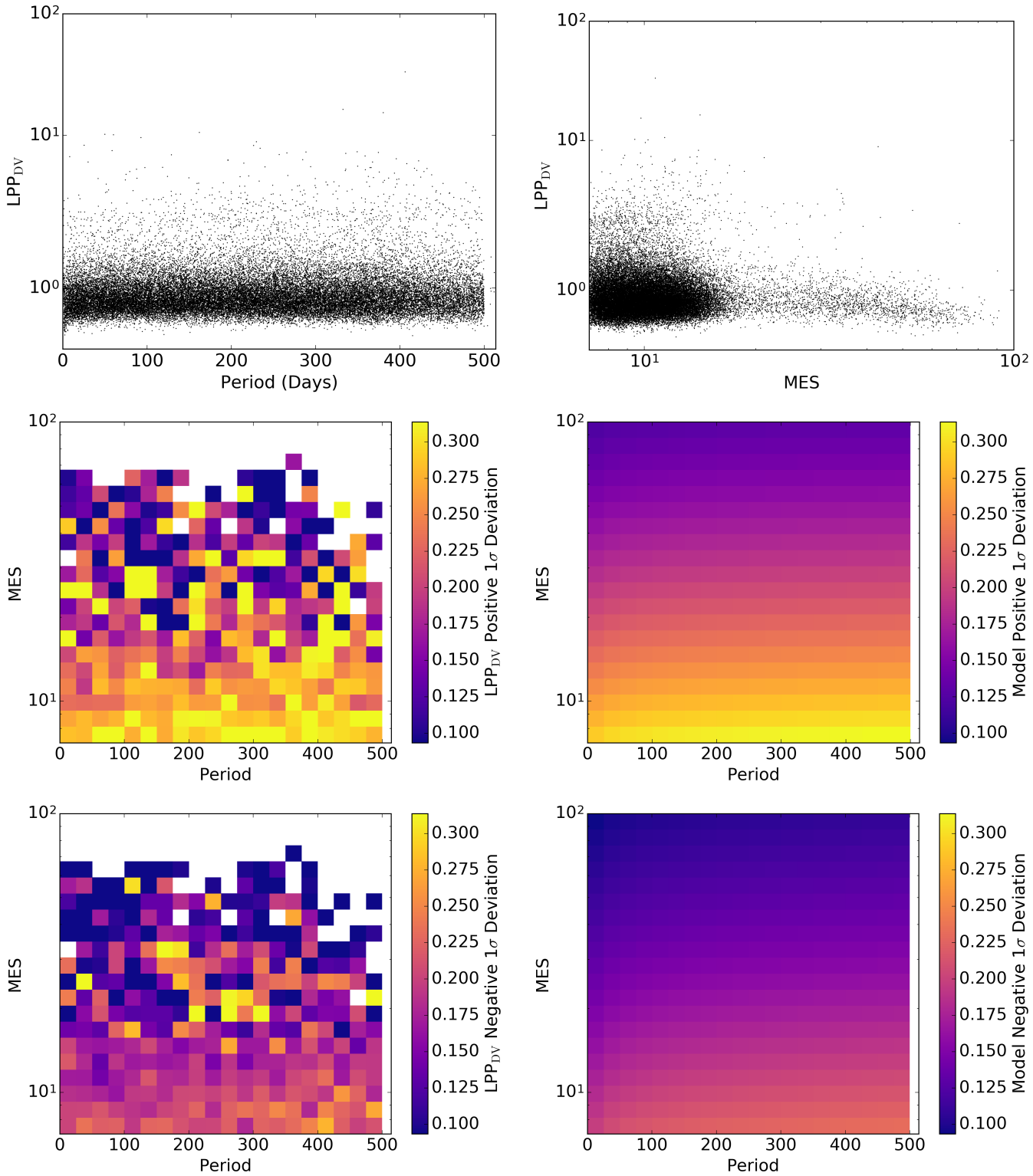

Figure 5. The top-left plot shows the LPP_{DV} values of all on-target injected planets on FGK dwarf targets as a function of period, and the top-right shows them as a function of MES. The middle-left plots shows the measured positive 1σ deviation (in the same units as LPP_{DV}) as a function of MES and period, and the middle-right plot shows the resulting best-fit model. The bottom plots show the same thing, but for the negative 1σ deviation (again in the same units as LPP_{DV}). These resulting model distributions are used when computing the Robovetter disposition score.

to calculate the reliability of the catalog from our false alarm sets is to first calculate how often the Robovetter correctly identifies false alarms as FPs, a value we call effectiveness (E). Then, given the number of FPs we identify in the obsTCE set, we determine the reliability of the catalog against the type of false alarms present in the simulated sets (invTCEs and scrTCEs). This method assumes the relative frequency of the different types of false alarms is well emulated by the simulated data sets, but does not require the total number of false alarms to be well emulated.

Robovetter effectiveness (E) is defined as the fraction of FPs correctly identified as FPs in the obsTCE data set,

$$E \equiv \frac{N_{\mathrm{FP}_{\mathrm{obs}}}}{T_{\mathrm{FP}_{\mathrm{obs}}}} \tag{3}$$

where $T_{\mathrm{FP}_{\mathrm{obs}}}$ is the number of identified FPs which are truly FPs and $N_{\mathrm{FP}_{\mathrm{obs}}}$ is the total number of measured FPs. Notice we are using N to indicate the measured number, and T to indicate the "True" number.

If the simulated (sim) false alarm TCEs accurately reflect the obsTCE false alarms, E can be estimated as the number of simulated false alarm TCEs dispositioned as FPs ($\mathrm{N}_{\mathrm{FP}_{\mathrm{sim}}}$) divided by the number of simulated TCEs ($\mathrm{N}_{\mathrm{sim}}$),

$$E \approx \frac{N_{\mathrm{FP}_{\mathrm{sim}}}}{N_{\mathrm{sim}}} \tag{4}$$

For our analysis of the DR25 catalog, we primarily use the union of the invTCEs and the scrTCEs as the population of simulated false alarms when calculating E, see §7.3.

Recall that the Robovetter makes a binary decision, and TCEs are either PCs or FPs. For this derivation we do not take into consideration the reason the Robovetter calls a TCE an FP (i.e., some false alarms fail because the Robovetter indicates there is a stellar eclipse or centroid offset). For most of parameter space, an overwhelming fraction of FPs are false alarms in the obsTCE data set. Future studies will look into separating out the effectiveness for different types of FPs using the set of injected astrophysical FPs (see §2.3).

At this point we drop the *obs* and *sim* designations in subsequent equations, as the simulated false alarm quantities are all used to calculate E. The N values shown below refer entirely to the number of PCs or FPs in the obsTCE set so that $N = N_{\mathrm{PC}} + N_{\mathrm{FP}} = T_{\mathrm{PC}} + T_{\mathrm{FP}}$. We rewrite the definition for reliability (Eq. 2) in terms of the number of true false alarms in obsTCE, T_{FP},

$$R \equiv \frac{T_{\mathrm{PC}}}{N_{\mathrm{PC}}} = 1 + \frac{T_{\mathrm{PC}} - N_{\mathrm{PC}}}{N_{\mathrm{PC}}} = 1 + \frac{N - T_{\mathrm{FP}} - N_{\mathrm{PC}}}{N_{\mathrm{PC}}} \tag{5}$$

When we substitute $N_{\mathrm{FP}} = N - N_{\mathrm{PC}}$ in Equation 5 we get another useful way to think about reliability, as one minus the number of unidentified FPs relative to the number of candidates,

$$R = 1 - \frac{T_{\mathrm{FP}} - N_{\mathrm{FP}}}{N_{\mathrm{PC}}} \tag{6}$$

However, the true number of false alarms in the obsTCE data set, T_{FP}, is not known. Using the effectiveness value (Equation 4) and combining it with our definition for effectiveness (Equation 3) we get,

$$T_{FP} = \frac{N_{\mathrm{FP}}}{E} \tag{7}$$

and substituting into equation 6 we get,

$$R = 1 - \frac{N_{\mathrm{FP}}}{N_{\mathrm{PC}}} \left(\frac{1 - E}{E} \right) \tag{8}$$

which relies on the approximation of E from Equation 4 and is thus a measure of the catalog reliability using all measurable quantities.

This method to calculate reliability depends sensitively on the measured effectiveness which relies on how well the set of known false alarms match the false alarms in the obsTCE data set. For example, a negative reliability can result if the measured effectiveness is lower than the true value. In these cases, it implies that there should be more PCs than exist, i.e., the number of unidentified false alarms is smaller than the number of remaining PCs to draw from.

4.2. *The Similarity of the Simulated False Alarms*

In order to use the scrTCE and invTCE sets to determine the reliability of our catalog we must assume that the properties of these simulated false alarms are similar to those of the false alarms in the obsTCE set. Specifically, this simulated data should mimic the observed not transit-like FPs, e.g., instrumental noise and stellar spots. For instance, our assumptions break down if all of the simulated false alarms were long-duration rolling-band FPs, but only a small fraction of the observed FPs were caused by this mechanism. Stated another way, the method we use to measure reliability, hinges on the assumption that for a certain parameter space the fraction of a particular type of FP TCEs is the same between the simulated and observed data sets. This is the reason we removed the TCEs caused by KOIs and eclipsing binaries in the simulated data sets (see §2.3.3). Inverted eclipsing binaries and transits are not the type of FP found in the obsTCE data set. Since the Robovetter is very good at eliminating inverted transits, if they were included, we would have an inflated value for the effectiveness, and thus incorrectly measure a higher reliability.

Figure 2 demonstrates that the number of TCEs from inversion and scrambling individually is smaller than the number of obsTCEs. At periods less than ≈100 days

this difference is dominated by the lack of planets and eclipsing binaries in the simulated false alarm data sets. At longer periods, where the TCEs appear to be dominated by false alarms, this difference is dominated by the cleaning (§2.3.3). Effectively, we search a significantly smaller number of stars for instances of false alarms. The deficit is also caused by the fact that all types of false alarms are not accounted for in these simulations. For instance, the invTCE set will not reproduce false alarms caused by sudden dropouts in pixel sensitivity caused by cosmic rays (i.e., SPSDs). The scrTCE set will not reproduce the image artifacts from rolling band because the artifacts are not as likely to line-up at exactly one *Kepler*-year. However, despite these complications, the period distribution of false alarms in these simulated data sets basically resembles the same period distribution as the obsTCE FP population once the two simulated data sets are combined. And since reliability is calculated using the fraction of false alarms that are identified (effectiveness), the overabundance that results from combining the sets is not a problem.

Another way to judge how well the simulated data sets match the type of FP in the obsTCEs is to look at some of the Robovetter metrics. Each metric measures some aspect of the TCEs. For example, the LPP Metric measures whether the folded and binned light curves are transit shaped, and Skye measures whether the individual transits are likely due to rolling band noise. If the simulated TCEs can be used to measure reliability in the way described above, then the fraction of false alarms in any period bin caused by any particular metric should match between the two sets. In Figure 6 we show that this is basically true for both invTCEs and scrTCEs, especially for periods longer than 100 days or MES less than 15. Keep in mind that more than one metric can fail any particular TCE, so the sum of the fractions across all metrics will be greater than one. The deviations between TCE sets is as large as 40% for certain period ranges and such differences may cause systematic errors in our measurements of reliability. But, since the types of FPs overlap, it is not clear how to propagate this information into a formal systematic error bar on the reliability.

For our discussion of the reliability estimate, we are cautiously satisfied with this basic agreement. Given that neither of the two sets perform better across all regions of parameter space, and since having more simulated false alarms improves the precision on effectiveness, we have calculated the catalog reliability using a union of the scrTCE and invTCE sets after they have been cleaned as described in §2.3.3.

5. TUNING THE ROBOVETTER FOR HIGH COMPLETENESS AND RELIABILITY

As described in the previous section, the Robovetter makes decisions regarding which TCEs are FPs and PCs based on a collection of metrics and thresholds. For each metric we apply a threshold and if the TCE's metrics' values lies above (or below, depending on the metric) the threshold then the TCE is called a FP. Ideally the Robovetter thresholds would be tuned so that no true PCs are lost and all of the known FPs are removed; however, this is not a realistic goal. Instead we sacrifice a few injTCEs in order to improve our measured reliability.

How to set these thresholds is not obvious and the best value can vary depending on which population of planets you are studying. We used automated methods to search for those thresholds that passed the most injTCEs and failed the most invTCEs and scrTCEs. However, we only used the thresholds found from this automated optimization to inform how to choose the final set of thresholds. This is because the simulated TCEs do not entirely emulate the observed data and many of the metrics have a period and MES dependence. For example, the injections were heavily weighted towards long periods and low MES so our automated method sacrificed many of the short period candidates in order to keep more of the long period injTCEs. Others may wish to explore similar methods to optimize the thresholds and so we explain our efforts to do this below.

5.1. *Setting Metric Thresholds Through Optimization*

For the first step in Robovetter tuning, we perform an optimization that finds the metric thresholds that maximize the fraction of TCEs from the injTCE set that are classified as PCs (i.e., completeness) and minimizes the fraction of TCEs from the scrTCE and invTCE sets identified as PCs (minimizes ineffectiveness or $1 - E$.) Optimization varies the thresholds of the subset of Robovetter metrics described below, looking for those thresholds that maximize completeness and minimize ineffectiveness.

This optimization is performed jointly across a subset of the metrics described in §3. The set of metrics chosen for the joint optimization, called "optimized metrics" are: LPP (§A.3.1), the Model-shift uniqueness test metrics (MS_1, MS_2, and MS_3; §A.3.4), Max SES to MES (§A.3.5), and TCE Chases (§A.3.3). Both the DV and ALT versions of these metrics, when applicable, were used in the optimization.

Metrics not used in the joint optimization are incorporated by classifying TCEs as PCs or FPs using fixed *a priori* thresholds prior to optimizing the other metrics. After optimization, a TCE is classified as a PC only if it passes both the non-optimized metrics and the optimized metrics. Prior to optimization the fixed thresholds for these non-optimized metrics pass about 80% of the injTCE set, so the final optimized set can have at most 80% completeness. Note, the non-optimized metric thresholds for the DR25 catalog changed after doing these optimizations. The overall effect was that the final measured completeness of the catalog increased (see §7), especially for the low MES TCEs. If the opti-

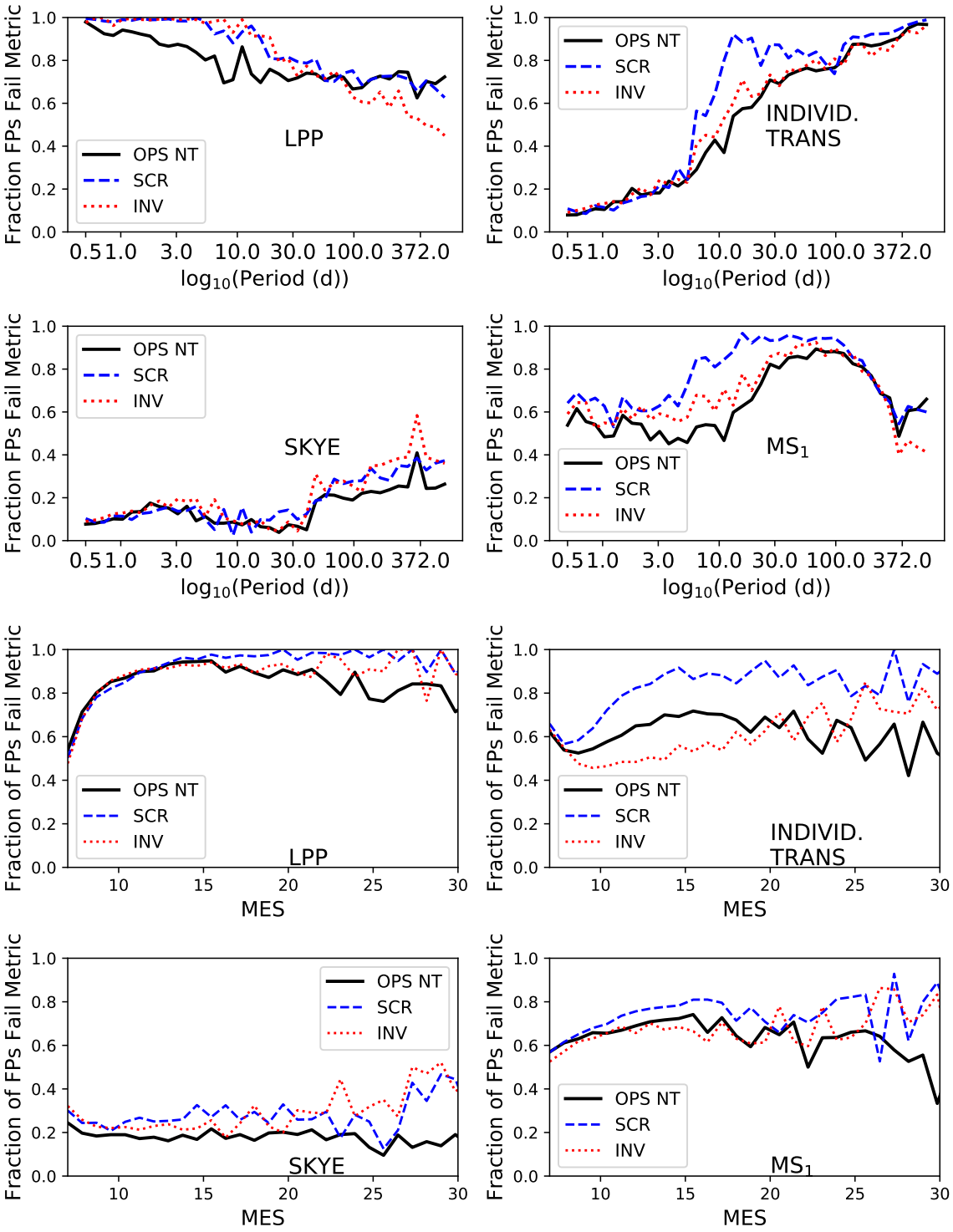

Figure 6. The fraction of not-transit-like FPs failed by a particular Robovetter metric plotted against the logarithm of the period (top two rows) or linear MES (bottom two rows). The fraction is plotted for the obsTCE set in black, the scrTCE set in blue, and the invTCE set in red. The metric under consideration is listed on each plot. For each metric we include fails from either detrending (DV or ALT). Upper left: LPP metric failures. Upper Right: TCEs that fail after removing a single transit due to any of the individual transit metrics. Lower left: TCEs that fail after removing a single transit due to the Skye metric. Lower right: Model Shift 1 metric failures. Notice that there is a basic similarity between the trends seen in the three data sets, especially at long periods and low MES.

mization were redone with these new thresholds, then it would find that the non-optimized metrics pass 90% of the injTCEs. We decided this change was not sufficient reason to rerun the optimization since it was only being used to inform and not set the final thresholds.

Optimization is performed by varying the selected thresholds, determining which TCEs are classified as PCs by both the optimized and non-optimized metrics using the new optimized thresholds, and computing C and $1 - E$. Our optimization seeks thresholds that minimize the objective function $\sqrt{(1 - \mathrm{E})^2 + (\mathrm{C} - \mathrm{C}_0)^2}$, where C_0 is the target completeness, so the optimization tries to get as close as possible to $1 - \mathrm{E} = 0$ and $\mathrm{C} = \mathrm{C}_0$. We varied C_0 in an effort to reduce the ineffectiveness. The thresholds are varied from random starting seed values, using the Nelder-Mead simplex algorithm via the MATLAB *fminsearch* function. This MATLAB function varies the thresholds until the objective function is minimized. There are many local minima, so the optimal thresholds depend sensitively on the random starting threshold values. The optimal thresholds we report are the smallest of 2000 iterations with different random seed values.

Our final optimal threshold used a target of $\mathrm{C}_0 = 0.8$, which resulted in thresholds that yielded $1 - \mathrm{E} = 0.0044$ and $\mathrm{C} = 0.799$. We experimented with smaller values of C_0, but these did not significantly lower ineffectiveness. We also performed an optimization that maximized reliability defined in §4.1 rather than minimizing ineffectiveness. This yielded similar results. We also explored the dependence of the optimal thresholds on the range of TCE MES and period. We found that the thresholds have a moderate dependence, while the ineffectiveness and completeness have significant dependence on MES and period range. Exploration of this dependence of Robovetter threshold on MES and period range is a topic for future study.

5.2. *Picking the Final Robovetter Metric Thresholds*

The results of this algorithmic optimization were used as a starting point for the final thresholds chosen for the DR25 catalog. We used the Confirmed Planet table and the Certified False Positive Table at the Exoplanet Archive, as well as the results of some prominent KOIs, to manually adjust the thresholds. Because most of the injTCEs, invTCEs, and scrTCEs are at long periods and low MES the automated tuning optimized the completeness and effectiveness for this part of the catalog. However, many of *Kepler*'s PCs have short periods and high SNR. The final catalog thresholds balanced the needs of the different parts of the catalog and endeavoured to keep the completeness of the long period candidates above 70%.

For those interested in a certain part of the KOI catalog, it may be better to re-tune the thresholds to optimize for higher reliability or to more aggressively remove certain types of false alarms. The Robovetter code[13] (and the Robovetter input files) are provided with the tunable thresholds listed at the top of the code. As an example, we include Table 4 as a list of the easily tunable thresholds for the metrics that determine whether an object is not transit-like. The table lists the thresholds we settled on for the DR25 catalog here, but it also provides the metrics for a higher reliability (lower completeness) catalog and a higher completeness (lower reliability) catalog. (These two different sets of thresholds are also included as commented-out lines in the Robovetter code after the set of thresholds used to create the DR25 catalog.) Each metric has its own range of possible values and some are more sensitive to small adjustments than others. Users should use caution when changing the thresholds and should endeavour to understand the different metrics, described in §3 and Appendix A, before doing so.

6. ASSEMBLING THE DR25 KOI CATALOG

The KOI catalog contains all the obsTCEs that the Robovetter found to have some chance of being transit-shaped, i.e., astrophysically transiting or eclipsing systems. All of the DR25 KOIs are fit with a transit model and uncertainties for each model parameter are calculated with a MCMC algorithm. We describe here how we decide which obsTCEs become KOIs, how we match the obsTCEs with previously known KOIs, and how the transit fits are performed. The KOI catalog is available in its entirety at the NASA Exoplanet Archive as the Q1–Q17 DR25 KOI Table[14].

6.1. *Creating KOIs*

The Robovetter gives every obsTCE a disposition, a reason for the disposition, and a disposition score. However, only those that are transit-like, i.e., have some possibility of being a transiting exoplanet or eclipsing binary system, are intended to be placed in the KOI catalog. For scheduling reasons, we created the majority of KOIs before we completed the Robovetter, so some not transit-like KOIs have been included in the KOI catalog. Using the final set of Robovetter dispositions, we made sure to include the following obsTCEs in the KOI table: 1) those that are "transit-like", i.e., are not marked with the NT-flag, and 2) KOIs that are not transit-like FPs which have a score value larger than 0.1. This last group were included to ensure that obsTCEs that marginally failed one Robovetter metric were easily accessible via the KOI catalog and given full transit fits with MCMC error bars. As in previous catalogs, all DR25 obsTCEs that federate (§6.2) to a previously identified KOI are included in the DR25 KOI table even if the Robovetter

[13] https://github.com/nasa/kepler-robovetter

[14] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr25_koi

Table 4. Robovetter Metric Thresholds

Test Name	Variable Name	DR25	High C	High R
SWEET	SWEET_THRESH	50	50	50
Max SES to MES	SES_TO_MES_THRESH	0.8	0.9	0.75
TCE CHASES	ALL_TRAN_CHASES_THRESH	0.8	1.0	0.55
MS_1 DV	MOD_VAL1_DV_THRESH	1.0	2.4	-1.0
MS_2 DV	MOD_VAL2_DV_THRESH	2.0	5.0	-0.7
MS_3 DV	MOD_VAL2_DV_THRESH	4.0	7.5	-1.6
MS_1 ALT	MOD_VAL1_ALT_THRESH	-3.0	-2.5	-4.3
MS_2 ALT	MOD_VAL2_ALT_THRESH	1.0	-0.5	2.5
MS_3 ALT	MOD_VAL3_ALT_THRESH	1.0	0.5	0.2
LPP DV	LPP_DV_THRESH	2.2	2.8	2.7
LPP ALT	LPP_ALT_THRESH	3.2	3.2	3.2

set the disposition to a not transit-like FP. All previous KOIs that were not found by the DR25 *Kepler* Pipeline (i.e., did not trigger a DR25 obsTCE) are not included in the DR25 KOI table at the Exoplanet Archive.

6.2. *Federating to known KOIs*

All obsTCEs that were included in the KOI catalog were either federated to known KOIs or given a new KOI number. Since KOIs have been identified before, federating the known KOIs to the TCE list is a necessary step to ensure that we do not create new KOIs out of events previously identified by the *Kepler* pipeline. The process has not changed from the DR24 KOI catalog (see §4.2 of Coughlin et al. 2016), so we simply summarize it here. For each obsTCE we use the ephemeris to determine what fraction of in-transit cadences overlap with all known KOIs on the same star. Those with significant overlap are considered federated. Also, those that are found at double or half the period are also considered matches (244 KOIs in total).

In some cases our automated tools want to create a new KOI in a system where one of the other previously known KOIs in the system did not federate to a DR25 TCE. In these cases we inspect the new system by hand and ensure that a new KOI number is truly warranted. If it is, we create a new KOI. If not, we ban the event from the KOI list. For instance, events that are caused by video cross-talk (Van Cleve & Caldwell 2016) can cause short-period transit events to appear in only one quarter each year. As a result, the *Kepler* Pipeline finds several one-year period events for an astrophysical event that is truly closer to a few days. In these cases we federate the one found that most closely matches the known KOI and we ban any other obsTCEs from creating a new KOI around this star. In Table 5 we report the entire list of obsTCEs that were not made into KOIs despite being dispositioned as transit-like (or not transit-like with a disposition score ≥ 0.1) and the automated federation telling us that one was appropriate. To identify the TCEs we specify the Kepler Input Catalog number and the planet number given by the *Kepler* Pipeline (Twicken et al. 2016).

Table 5. obsTCEs Banned from Becoming KOIs (§6.2)

TCE-ID (KIC-PN)
003340070-04
003958301-01
005114623-01
005125196-01
005125196-02
005125196-04
005446285-03
006677841-04
006964043-01
006964043-05
007024511-01
008009496-01
008956706-01
008956706-06
009032900-01
009301564-01
010223616-01
012459725-01
012644769-03

It is worth pointing out that the banned TCEs is the one pseudo-manual step that is not repeated for all the simulated TCEs. These banned TCEs effectively disappear when doing statistics on the catalog (i.e., these TCEs do not count as either a PC or an FP.) They are not present in the simulated data sets, nor are we likely to remove good PCs from our sample this way. Most banned TCEs are caused by either a short-period binary whose flux is contaminating our target star (at varying depths through mechanisms such as video cross-talk or reflection), or are systems with strong TTVs (transit timing variations, see §6.3). In both cases, the Pipeline finds several TCEs at various periods, but only one astrophysical system causes the signal. By banning these obsTCEs, we are removing duplicates from the KOI catalog and improving the completeness and reliability statistics reported in §7.3.

6.3. *KOI Transit Model Fits*

Each KOI, whether from a previous catalog or new to the DR25 catalog, was fit with a transit model in a consistent manner. The model fitting was performed in a similar to that described in §5 of Rowe et al. (2015a). The model fits start by detrending the DR25 Q1–Q17 PDC photometry from MAST[15] using a polynomial filter as described in §4 of Rowe et al. (2014). A transit model based on Mandel & Agol (2002) is fit to the photometry using a Levenberg-Marquardt routine (More et al. 1980) assuming circular orbits and using fixed quadratic limb darkening coefficients (Claret & Bloemen 2011) calculated using the DR25 stellar parameters (Mathur et al. 2017). TTVs are included in the model fit when necessary; the calculation of TTVs follows the procedure described in §4.2 of Rowe et al. (2014). The 296 KOIs with TTVs and the TTV measurements of each transit are listed in Table 6. The uncertainties for the fitted parameters were calculated using a Markov-chain Monte Carlo (MCMC) method (Ford 2005) with a single chain with a length of $2{\times}10^5$ calculated for each fit. In order to calculate the posterior distribution the first 20% of each chain was discarded. The transit model fit parameters were combined with the DR25 stellar parameters and associated errors (Mathur et al. 2017) in order to produce the reported planetary parameters and associated errors. The MCMC chains are all available at the Exoplanet Archive and are documented in Hoffman & Rowe (2017).

The listed planet parameters come from the least-squares (LS) model fits and the associated errors from the MCMC calculations. Note that not all KOIs could be successfully modelled, resulting in three different fit types: LS+MCMC, LS, and none. In the case of LS+MCMC the KOIs were fully modelled with both a least-squares model fit and the MCMC calculations were completed to provide associated errors. In the cases where the MCMC calculations did not converge, but there is a model fit, the least-square parameters are available without uncertainties (LS fit type). In the final case, where a KOI could not be modelled (e.g., cases where the transit event was not found in the detrending used by the MCMC fits) only the period, epoch, and duration of the federated TCE are reported and the fit type is listed as none.

[15] https://archive.stsci.edu/kepler/

Table 6. TTV Measurements of KOIs

n	t_n	TTV_n	$TTV_{n\sigma}$
	BJD-2454900.0	days	days
KOI-6.01			
1	54.6961006	0.0774247	0.0147653
2	56.0302021	-0.0029102	0.0187065
3	57.3643036	-0.0734907	0.0190672
4	58.6984051	0.0119630	0.0176231
...	...	...	...
KOI-8.01			
1	54.7046603	-0.0001052	0.0101507
2	55.8648130	-0.0103412	0.0084821
3	57.0249656	0.0047752	0.0071993
...	...	...	...
KOI-8151.01			
1	324.6953389	0.1093384	0.0025765
2	756.2139285	-0.3478332	0.0015206
3	1187.7325181	0.0110542	0.0016480
...	...	...	...

NOTE—Column 1, n, is the transit number. Column 2, t_n, is the transit time in Barycentric Julian Date minus the offset 2454900.0. Column 3, TTV_n, is the observed - calculated (O-C) transit time. Column 4, $TTV_{n\sigma}$, is the 1σ uncertainty in the O-C transit time. Table 6 is published in its entirety in the electronic edition of the *Astrophysical Journal Supplement*. A portion is shown here for guidance regarding its form and content.

7. ANALYSIS OF THE DR25 KOI CATALOG

7.1. *Summary of the KOI Catalog*

The final DR25 KOI catalog, available at the NASA Exoplanet Archive, contains all TCEs that pass the not transit-like tests (§A.3) and those that fail as not transit-like with a disposition score ≥ 0.1. Some overall statistics of the DR25 KOI catalog are as follows:

- 8054 KOIs
- 4034 PCs
- 738 new KOIs
- 219 new PCs
- 85.2% of injTCEs are PCs
- 99.6% of invTCEs and scrTCEs are FPs

A plot of the planetary periods and radii is shown in Figure 7, with the color indicating the disposition score. The distribution of the periods and planetary radii of the planet candidates in this catalog is shown along the x- and y-axis. A clear excess of candidates exists with periods near 370 d; this excess disappears if we only consider those with a disposition score > 0.7. While the disposition score provides an easy way to make an additional cut on the PC population at long periods, when discussing the catalog PCs below we are using the pure dispositions of the Robovetter unless otherwise stated. The slight deficit of planets with radii just below $2.0\,R_\oplus$ is consistent with the study of Fulton et al. (2017) where they report a natural gap in the abundance of planets between super-Earths and mini-Neptunes by applying precise stellar parameters to a subset of the *Kepler* transiting candidates (Johnson et al. 2017; Petigura et al. 2017b). The new KOIs with a disposition of PC are found at all periods, but only ten have MES ≥ 10.

7.2. *Comparison of Dispositions to Other Catalogs*

We compare the DR25 KOI catalog to two sets of *Kepler* exoplanets: the confirmed exoplanets and the certified FPs. In both of these cases, additional observations and careful vetting are used to verify the signal as either a confirmed exoplanet or a certified FP (Bryson et al. 2017). It is worth comparing the Robovetter to these catalogs as a sanity check.

We use the confirmed exoplanet list from the Exoplanet Archive[16] on 2017-05-24. 2279 confirmed planets are in the DR25 KOI catalog. The DR25 Robovetter fails 44 of these confirmed planets, or less than 2%. Half of these FPs are not transit-like fails, 16 are stellar eclipse fails, six are centroid offsets, and one is an ephemeris match. Twelve fail due to the LPP metric; all of these twelve have periods less than 50 days. The LPP metric threshold was set to improve the reliability of the long period KOIs, an act which sacrificed some of the short period KOIs. The reason the Robovetter failed each confirmed planet is given in the "koi_comment" column at the Exoplanet Archive (see §B).

For the vast majority of the Robovetter FPs on the confirmed planet list, careful inspection reveals that there is no doubt that the Robovetter disposition is incorrect. As an example, Kepler-10b (Batalha et al. 2011; Fogtmann-Schulz et al. 2014), a rocky planet in a 0.84 d orbit, was failed due to the LPP metric. This occurred because the detrending algorithm (the harmonic identification and removal algorithm in TPS, see Jenkins 2017b) used by the *Kepler* Pipeline significantly distorts the shape of the transit, a known problem for strong, short period signals (Christiansen 2015). The LPP metric, which compares the shape of the folded light curve to known transits, then fails the TCE.

In some cases the Robovetter may be correctly failing the confirmed planet. Many of the confirmed planets are only statistically validated (Morton et al. 2016; Rowe et al. 2014). In these cases no additional data exists proving the physical existence of a planet outside of the transits observed by *Kepler*. It is possible that the DR25 light curves and metrics have now revealed evidence that the periodic events are caused by noise or a binary star. For example, Kepler-367c (Rowe et al. 2014), Kepler-1507b (Morton et al. 2016), and Kepler-1561b (Morton et al. 2016) (KOIs 2173.02, 3465.01, and 4169.01, respectively) were all confirmed by validation and have now failed the Robovetter because of the new ghost metric (see §A.5.2), indicating that the events are caused by a contaminating source not localized to the target star. These validations should be revisited in the light of these new results.

It is also worth noting that none of the confirmed circumbinary planets (e.g., Doyle et al. 2011; Orosz et al. 2012) are in the DR25 KOI catalog. However, the eclipsing binary stars that they orbit are listed as FPs. The timing and shape of the circumbinary planet transits vary in a complicated manner, making them unsuitable for detection by the search algorithm used by the *Kepler* Pipeline to generate the DR25 obsTCE list. As a result, this catalog cannot be used for occurrence rate estimates of circumbinary planets, and their absence in the KOI catalog should not cast doubt on their veracity.

We use the Certified False Positive table[17] downloaded from the Exoplanet Archive on 2017-07-11 to evaluate the performance of the Robovetter at removing known FPs. This table contains objects known to be FPs based on all available data, including ground-based follow-up information. The Robovetter passes 106 of the 2713 certified FPs known at the time, only 3.9 per cent. Most of those called PCs by the Robovetter are high signal-to-noise and more than half have periods less than 5 days. The most common reason they are certified FPs is that there is evidence they are eclipsing binaries. In some cases, external information, like radial velocity measurements, provide a mass which determines that the KOI is actually a binary system. The other main reason for the discrepancy between the tables is that the certified FPs often show evidence of significant centroid offsets. In crowded fields the Centroid Robovetter (§A.5.1) will not fail observed offsets because of the potential for confusion. For the Certified False Positive table, individual cases are examined by a team of scientists who determines when there is sufficient proof that the signal is indeed caused by a background eclipsing binary.

[16] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets

[17] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=fpwg

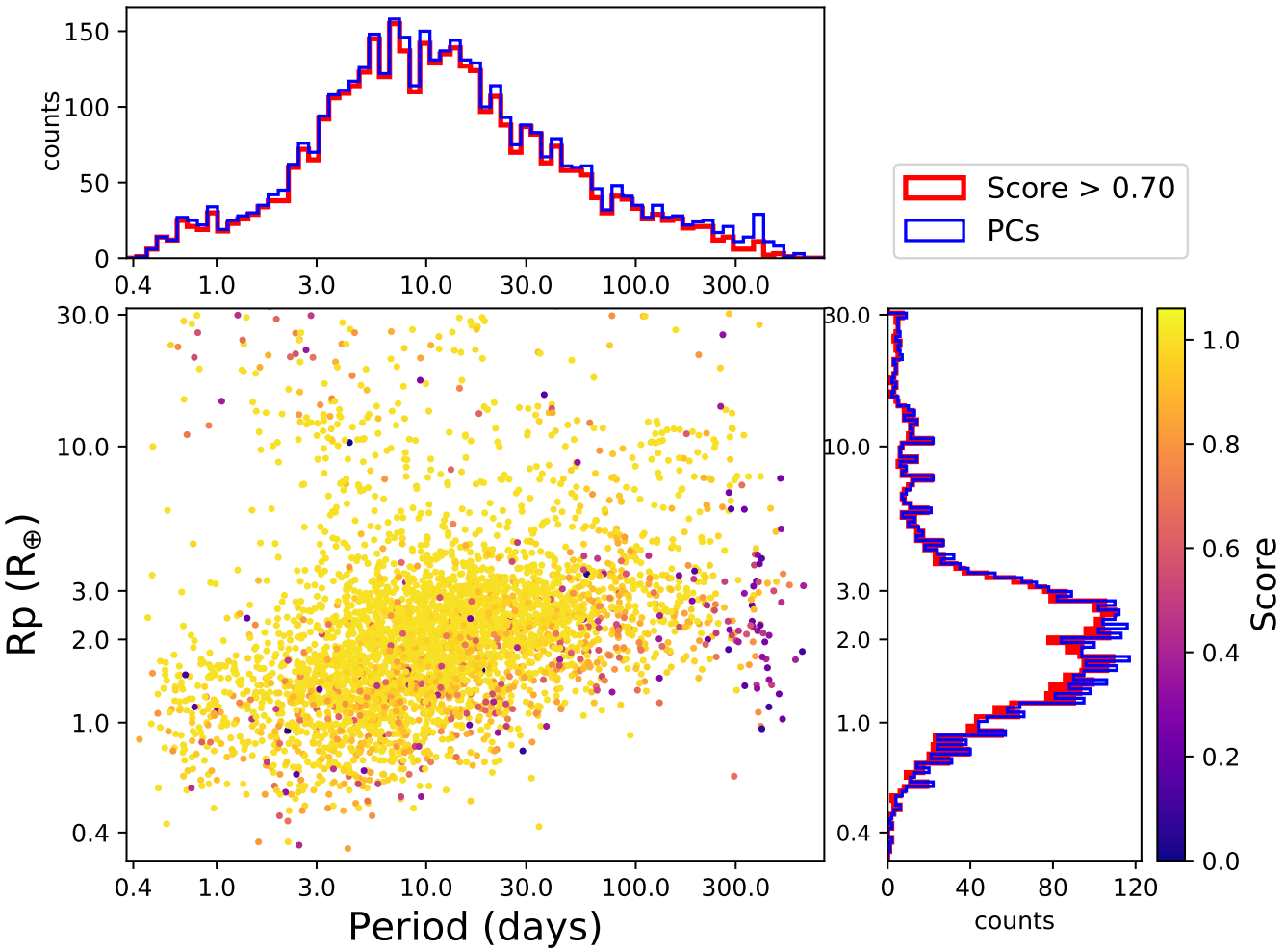

Figure 7. DR25 PCs plotted as planet radius versus period with the color representing the disposition score. The period and planet radii distributions are plotted on the top and on the left, respectively, in blue. The red line shows the distributions of those PCs with a disposition score greater than 0.7. The excess of PCs at long-periods disappears when cutting the population on disposition score.

7.3. *Catalog Completeness, Effectiveness, and Reliability*

To evaluate the performance of the Robovetter and to measure the catalog completeness and reliability, we run the Robovetter on the injTCEs, invTCEs, and scrTCEs. As a high level summary, Figure 8 provides the completeness, in-effectiveness ($1-E$), and reliability for a 3 by 3 grid across period and MES. If the same figure is made for only the FGK dwarf type stars ($\log g \geq 4.0$ and 4000 K $\geq T_\star <$ 7000 K), the long period, low MES bin improves substantially. Giant stars are inherently noisy on time scales of planet transits (see Figure 9 of Christiansen et al. 2012) causing more FPs and also causing more real transits to be distorted by the noise. For FGK dwarf stars and only considering candidates with periods between 200 d and 500 d and MES < 10, $C = 76.7\%$, $1-E = 1.1\%$, and $R = 50.3\%$, which is a 13.1 percentage point improvement in reliability and 3 percentage point improvement in completeness compared to all stars in the same period and MES range.

7.3.1. *Completeness*

The completeness of the vetting is measured as the fraction of injTCEs that are dispositioned as PCs. We discuss here the detection efficiency of the Robovetter, not the Kepler Pipeline (see §8 for a discussion of the Pipeline completeness). Across the entire set of recovered injTCEs which have periods ranging from 0.5–500 d, the Robovetter dispositioned 85.2% as PC. As expected, the vetting completeness is higher for transits at shorter periods and higher MES, and lower for longer periods and lower MES. The right hand column of Figure 9 shows how the completeness varies with period, expected MES, number of transits, and transit duration. Note that expected MES is the average MES at which the injected transit signal would be measured in

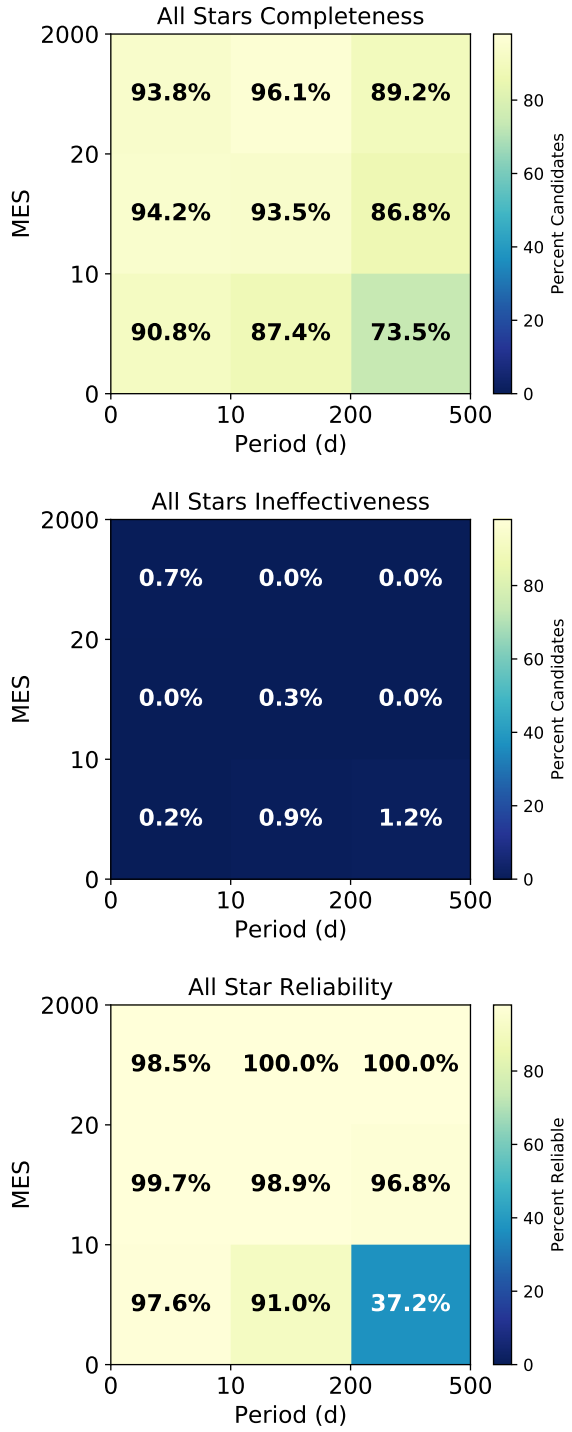

Figure 8. A coarse binning of the completeness (C), ineffectiveness (1-E), and reliability (R) for different period and MES bins (shown from top to bottom, respectively). The effectiveness and reliability are based on the combined invTCE and scrTCE data sets. Notice that the Robovetter effectiveness at removing these false alarms is incredibly high, but for long periods and low MES the resulting reliability is lower because of the large number of false alarms and small number of true planets. For FGK dwarf stars only, the reliability is 50.3% and the completeness is 76.7% for planets in the longest period, lowest MES box.

the target light curve, given the average photometric noise of that light curve and the depth of the injected transit signal — see Christiansen 2017 for more details. The small drop in completeness just short of 90 days is likely caused by the odd-even metric (§A.4.1.4), which only operates out to 90 days, confusing true transits for binary eclipses.

Because most planet occurrence rate calculations are performed using period and radius (e.g., Burke et al. 2015), we show the measured completeness binned by period and radius in Figure 10. The plot is linear in period and radius in order to emphasize the long period planets. Planetary radius is not a natural unit to understand the performance of the Robovetter since it combines the depth of the transit, the noise level of the light curve, and the stellar radius. At the longest periods the Robovetter more often fails the largest planets, though the trend is reduced when only considering the FGK dwarf stars. The largest radii planets in the injTCE population are entirely around giant stars, because large planets were not injected onto the dwarf stars (Christiansen 2017). The giant stars are notoriously noisy. As a result the largest radii planets in the injTCEs are more likely to be dispositioned incorrectly. Also, even when only considering the dwarf stars, a larger fraction of the big planets will be around larger, more massive stars (in comparison to the small planets which will mostly be found around smaller stars). This results in a population of planets that produce longer transit durations. The Robovetter performs less well for long transit durations (see Figure 9). For more figures showing the Robovetter effectiveness across different parameters, see Coughlin (2017b).

7.3.2. *Effectiveness*

The effectiveness of the Robovetter at identifying instrumental and stellar noise is calculated using the union of the invTCEs and scrTCEs (see §4.1), after removing the TCEs specified in §2.3.3. Across the entire set, the Robovetter dispositions 99.6% of these simulated false alarms as FPs. Only 119 of the 28,735 simulated false alarms are dispositioned as a PC by the Robovetter. Most of these invPCs and scrPCs are at long periods and low MES. However, using the 4544 invTCEs and scrTCEs that have periods between 200 d and 500 d and MES less than 10, the Robovetter's effectiveness is 98.8% (see Figure 8). Unfortunately, because there are so few candidates at these long periods, this translates to a relatively low reliability. For detailed plots showing how effectiveness varies with different parameters see Coughlin (2017b).

7.3.3. *Reliability*

The reliability is measured according to the method described in §4.1 using the effectiveness measured from the combined scrTCE and invTCE data sets and the number of observed PCs. If one bins over the entire data

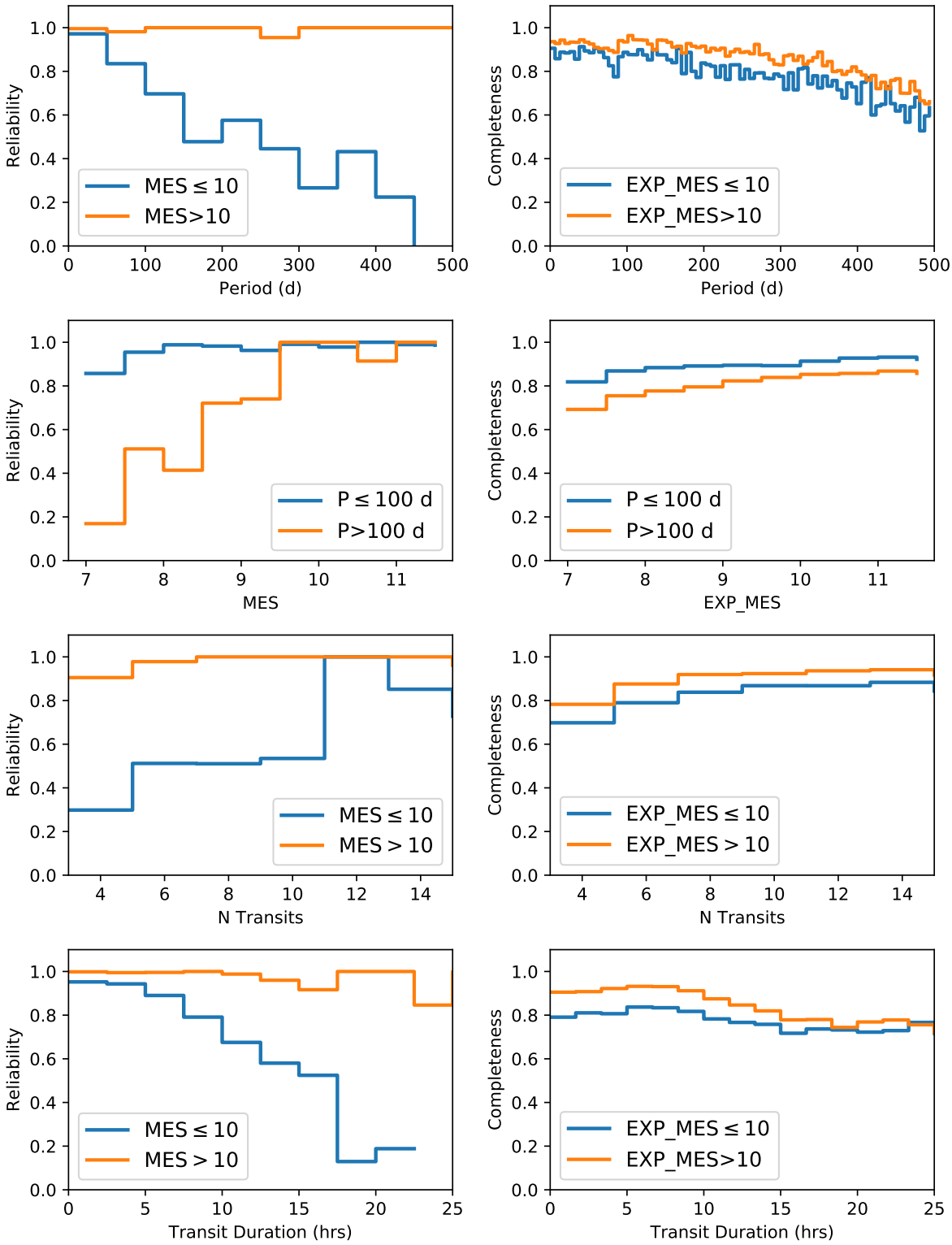

Figure 9. The reliability (left) and completeness (right) of the DR25 catalog plotted as a function of period, MES, number of transits, and transit duration. In each case the blue line is for those with MES $\leq$ 10 or periods $\leq$ 100 d. The orange line shows the completeness or reliability for the rest of the population (see the legend for each plot). EXP_MES is the expected MES (see Christiansen 2017 and §7.3.1).

set, the overall reliability of the catalog is 97%. However, as Figure 9 demonstrates, the reliability for long period, and especially low MES planets, is significantly smaller. For periods longer than 200 d and MES less than 10, the reliability of the catalog is approximately 37%, i.e., 6 out of 10 PCs are caused by noise. As with completeness, we plot the reliability as a function of period and planet radius in Figure 11. The least reliable planets are at long periods and have radii less than $4R_{\oplus}$.

The uncertainty in the reliability is likely dominated by how well the false alarms in the scrTCE and invTCE sets match the false alarms in the obsTCE data set (see §4.2 for further discussion on their similarity). One way to get a handle on the uncertainty on reliability is to calculate the reliability in three different ways for the long period (200–500 d), low MES (< 10) obsTCEs. First, we use only the invTCEs to measure the effectiveness at removing false alarms. This results in a lower reliability, namely $R = 24\%$ with $E = 98.5\%$. Second, we use only the scrTCEs to measure the effectiveness. This results in a higher reliability, $R = 51\%$ with $E = 99.1\%$. Third, we select, at random, half of the combined population of false alarms (scrTCE and invTCE) and calculate the reliability. After doing this random selection 100 times, we obtained $R = 38\%$ with a standard deviation of 8%, and the distribution appears symmetric and basically Gaussian in shape.

The Robovetter is less effective at removing the false alarms produced by inversion than those by scrambling the data. Inversion finds false alarms with periods near 372 d, which are frequently caused by image artifacts. Scrambling under-populates these types of false alarms, and since they are the difficult to eliminate, it is not surprising that the reliability measured by inversion is worse than scrambling. The truth likely lies somewhere in between. We encourage users of these data sets to consider ways to optimize the reliability measurement, and the error budget associated with them, when doing occurrence rate calculations.

7.3.4. *High Reliability Using the Disposition Score*

The disposition scores discussed in §3.2 can be used to select a more reliable, though more incomplete, sample of planet candidates. In Figure 12 we show the distribution of disposition scores for the PCs and FPs from the observed, inverted, scrambled, and on-target planet injection populations. (Note, the inverted and scrambled populations have been cleaned as discussed in §2.3.3). For all populations, the PC distribution tends to cluster near a score of 1.0 with a tail that extends towards lower score values. Lower MES values tend to have a greater proportion of lower score values. Similarly, the vast majority of FPs have a score of 0.0, with only a small fraction extending towards higher score values (note the y-axis for the FPs is logarithmic, while the y-axis for PCs is linear). Comparing the populations, the on-target planet injections have a greater concentration of score values towards 0.5 for both the PCs and FPs than other populations. Both the inverted and scrambled populations have very few PCs near high score values. We can exploit the relative distribution of PC and FP score values for the different populations to select a higher reliability catalog.

At the top of Figure 13 we show how the completeness and reliability of the catalog vary for different cuts on the disposition score for MES<10 and periods between 200 and 500 days. The effectiveness of the Robovetter increases as the score threshold is increased. The reliability values also depend on the number of observed PCs that remain, which is why reliability does not change in step with the effectiveness. Selecting the PC sample by choosing those with a disposition score above 0.6 (see the point labeled 0.6 on the top of Figure 13) yields an 85% reliability and a completeness that is still above 50%. Doing a score cut in this way not only removes those dispositioned as a PC from the sample, but also causes a few obsTCEs which are formally dispositioned as FPs to now be included in the sample. An FP with a high score occurs when a TCE marginally fails a single metric.

It is interesting to note that the number of inferred candidates, i.e., the number of candidates after accounting for the Robovetter completeness and catalog reliability, does not change significantly with the score cut. In the lower plot of Figure 13 we plot both the observed number of PCs and the corrected number of PCs that have periods between 200 and 500 days and MES less than 10. The correction is done by taking the number of PCs and multiplying by the reliability and dividing by the completeness. The error bars only include the Poisson counting error in the number of observed PCs and do not include errors in the measured completeness or reliability. The corrected number of PCs only varies by approximately 1σ regardless of the score cut used.

7.4. *Multiple-planet systems*

Lissauer et al. (2014) argues that almost all multi-planet, transit systems are real. Forty-seven, or 21%, of the new DR25 PCs are associated with targets with multiple PCs. One of the new PCs, KOI 82.06, is part of a six candidate system around the star *Kepler*-102. Five candidates have previously been confirmed (Marcy et al. 2014; Rowe et al. 2014) in this system. The new candidate is the largest radius confirmed planet in the system. It also lies a bit outside the 4:3 resonance; possibly adding to the excess of planets found just wide of such first-order resonances (Lissauer et al. 2011a). If verified, this would be only the 3rd system with six or more planets found by Kepler. The other new candidate within a high multiplicity system is KOI 2926.05. The other four candidates in this system around *Kepler*-1388 have been validated by Morton et al. (2016). This new candidate also orbits just exterior to a first-order mean

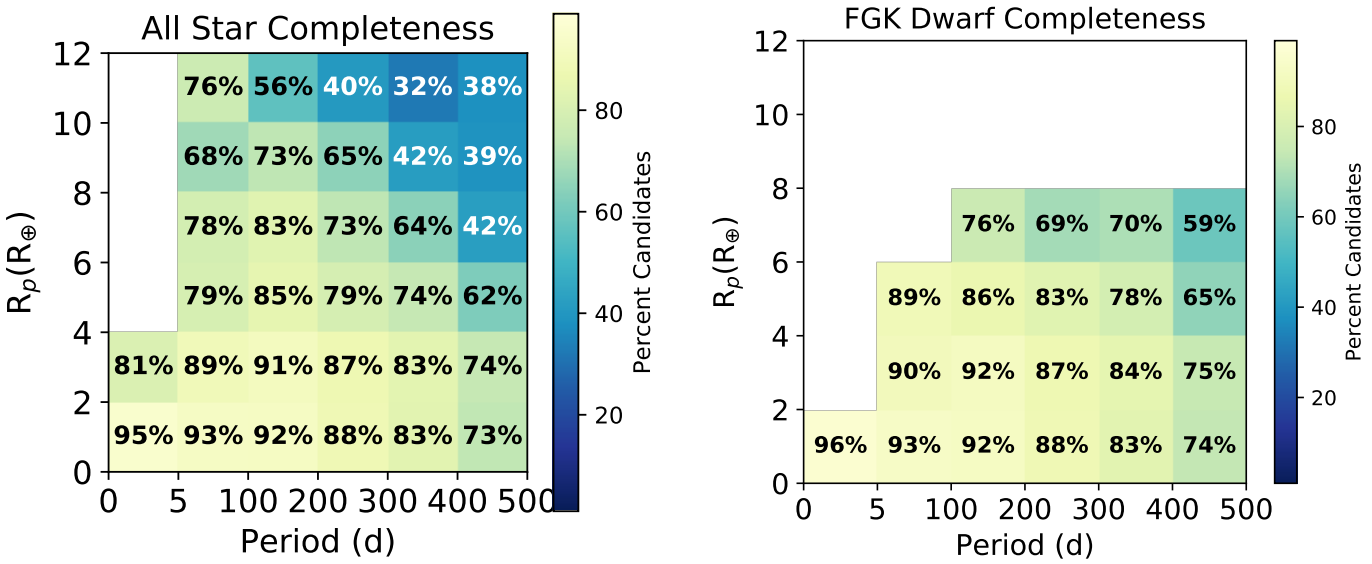

Figure 10. The Robovetter completeness binned by period and planet radius for all stars (left) and for only FGK dwarf stars (right). Bins with fewer than 10 injTCEs are not plotted.

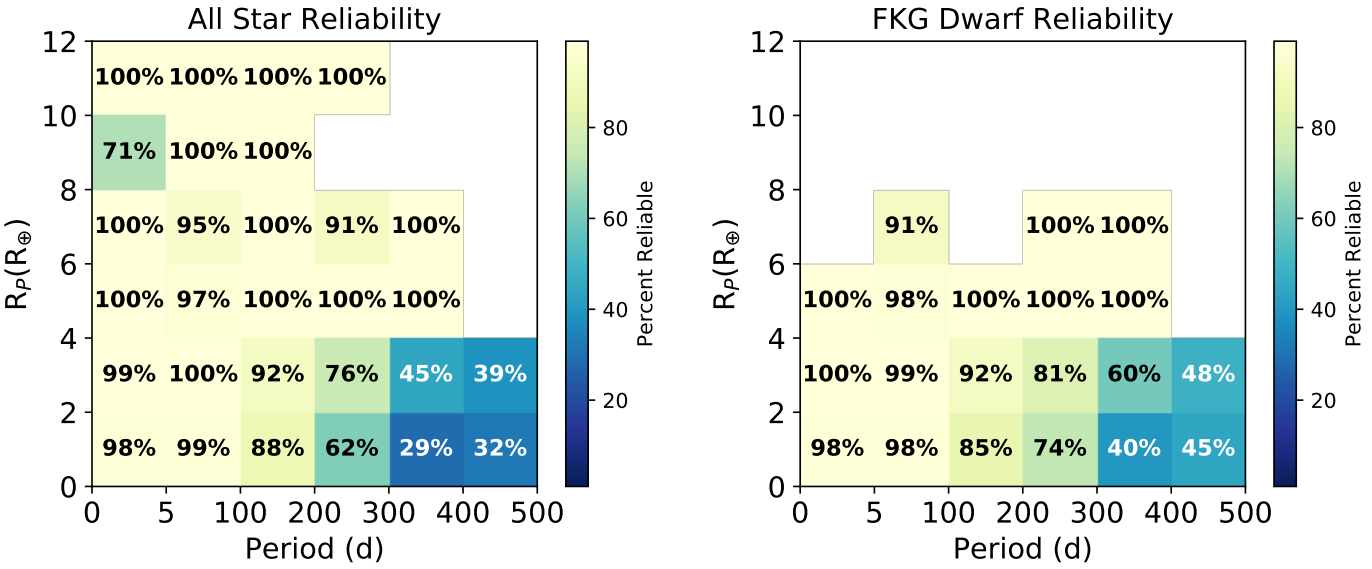

Figure 11. A 2D binning of the candidate catalog reliability for period and planet radius for all stars (left) and for the FGK dwarf stars (right). Bins with fewer than 3 candidates or fewer than 20 simulated false alarms (from invTCE and scrTCE) are not plotted.

motion resonance with one of the four previously known planets.

7.5. *Potentially Rocky Planets in the Habitable Zone*

Kepler is NASA's first mission capable of detecting Earth-size planets around Sun-like stars in one-year orbits. One of its primary science goals is to determine the occurrence rate of potentially habitable, terrestrial-size planets — a value often referred to as eta-Earth. Here we use the concept of a habitable zone to select a sample of planet candidates that are the right distance from their host stars and small enough to possibly have a rocky surface. A point that bears repeating is that no claims can be made regarding planetary habitability

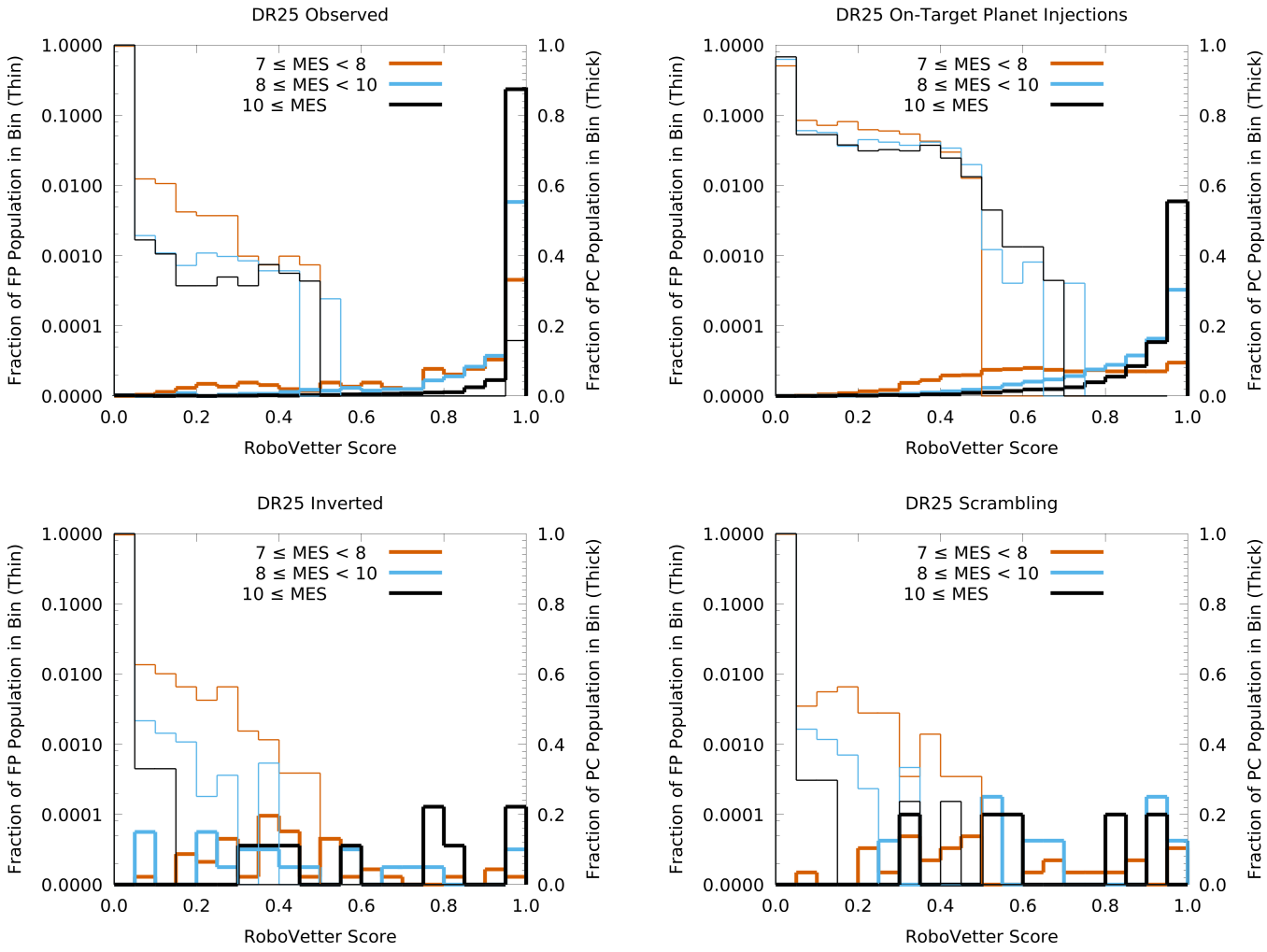

Figure 12. Plots of the score distribution of PCs (thick lines, right y-axis) and FPs (thin lines, left y-axis, logarithmic scaling) for the observed (top-left), on-target planet injections (top-right), inverted (bottom-left), and scrambled (bottom-right) TCEs.

based on size and orbital distance alone. This sample is, however, of great value to the occurrence rate studies that enable planet yield estimates for various designs of future life-detection missions (Stark et al. 2015). This eta-Earth sample is provided in Table 7 and shown in Figure 14.

7.5.1. *Selecting the Eta-Earth Sample*

Before applying thresholds on planet properties, we first select a sample based on disposition score (see §3.2) in order to produce a sample of higher reliablilty planets orbiting G-type stars. At long orbital period and small radius, we are vulnerable to instrumental false alarms despite the significant improvements afforded us by the latest versions of metrics like Marshall, Skye, Rubble, Chases, and Model-shift. This is evident in the FGK dwarf sample of Figures 10 and 11 by comparing the relatively low reliability (45%–74%) and completeness (74% to 88%) measurements in the bottom right boxes to others at shorter period and larger radius. Removing candidates with score < 0.5 results in a significant improvement in the sample reliability with a small degradation in the sample completeness (Figure 13). The candidates reported in Table 7 are ≈80% reliable for the G-type stars and even higher for the K- and M-type stars. Note, there is only one late F-type star in the sample. Kepler was not designed to reach the habitable zones of F-type stars, nor did the target list include many such stars.

The DR25 catalog uses the transit depth and period, along with the DR25 stellar table of Mathur et al. (2017), to derive the planet radius and the semi-major axis of the planet's orbit. From these we calculate the insolation flux in units of the Earth's insolation flux,

$$S_p = \frac{R_\star^2 \cdot (T_\star/5777)^4}{a^2}, \tag{9}$$

where a is the semi-major axis of the planet's orbit in AU, $T_\star$ is the host star temperature in Kelvin, 5777 K is

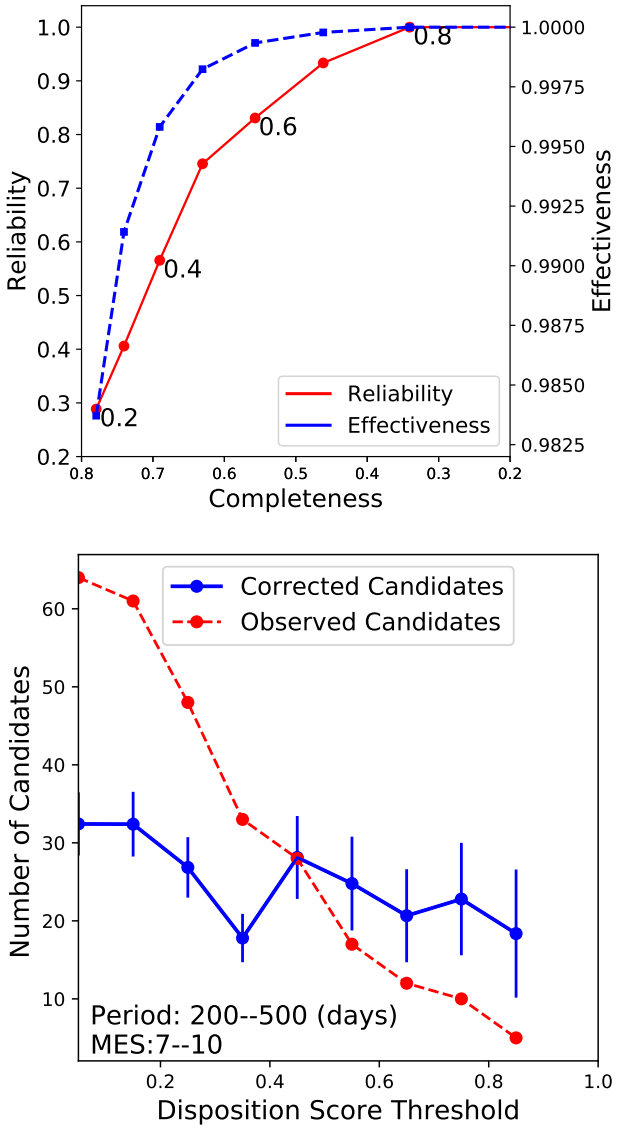

Figure 13. [Top] The reliability (red) and effectiveness (blue) of the DR25 catalog as a function of Completeness for MES $\leq$ 10 and periods between 200 and 500 d PCs that result when using different disposition score thresholds (shown as black numbers) to select the PCs. Higher disposition score thresholds result in higher reliability but lower completeness. Note, the completeness axis increases to the left. [Bottom] The number of PCs (in red) in the same period and MES space when making a cut on different disposition scores. The blue line corrects the number of candidates for the completeness and reliability. The error bars only reflect a Poisson error based on the number of observed planet candidates shown in red.

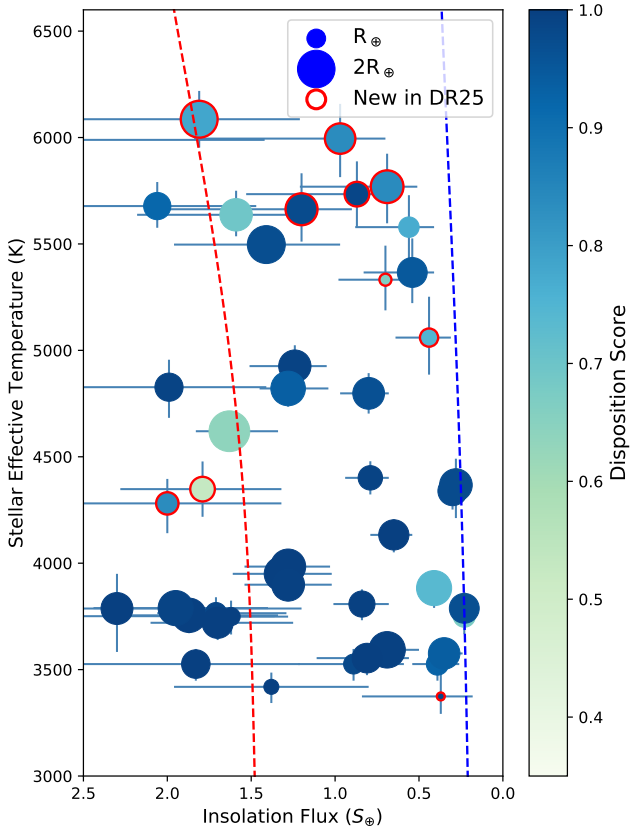

Figure 14. DR25, eta-Earth sample of PCs plotted as stellar effective temperature against insolation flux using the values reported in the DR25 KOI catalog (which uses stellar properties from the DR25 stellar catalog (Mathur et al. 2017). The size of the exoplanet is indicated by the size of the circle. The color indicates the disposition score. Only those with disposition score greater than 0.5 are plotted. Only objects whose error bars indicate that they could be in the habitable zone and have a radius less than 1.8 $R_\oplus$ are shown. Those with a red ring are new to the DR25 catalog.

the effective temperature of the Sun, $R_\star$ is the radius of the star in units of $R_\odot$, and thus S_p is in units relative to the Earth's insolation flux. The errors for both insolation flux and radii include the errors from the DR25 stellar catalog. The habitable zone represents a range of orbits where the flux received by the host star allows for the possibility of surface liquid water on an Earth-size planet. While the insolation limits for the habitable zone depends on the stellar temperature, it roughly falls from 0.2–1.7 $S_\oplus$ (see Figure 14). We use the empirical (recent Venus/early Mars) habitable zone of Kopparapu et al. (2013). To err on the side of inclusiveness, we include candidates whose one sigma error bars on the insolation flux overlap this empirical habitable zone.

Finally, we include only those candidates that satisfy the size constraint $R_p - \sigma_{R_p,\mathrm{low}} < 1.8\ R_\oplus$. The purpose of the size constraint is to identify candidates likely to have a bulk composition similar to terrestrial planets in the solar system. The 1.8 $R_\oplus$ upper limit is taken from Fulton et al. (2017) who report a distinct gap in the

Table 7. Habitable Zone Terrestrial-Sized Planet Candidates

KOI	KIC	Kepler	Period [days]	R_p [$R_\oplus$]	S_p [$S_\oplus$]	$T_\star$ [K]	$R_\star$ [$R_\odot$]	MES	Disp. Score
172.02	8692861	Kepler-69 c	242.46130	$1.73^{+0.21}_{-0.22}$	$1.59^{+0.59}_{-0.45}$	5637^{+113}_{-101}	$0.94^{+0.12}_{-0.12}$	18.0	0.693
238.03	7219825	···	362.97828	$1.96^{+0.33}_{-0.29}$	$1.81^{+0.87}_{-0.60}$	6086^{+133}_{-133}	$1.22^{+0.20}_{-0.18}$	11.9	0.784
438.02	12302530	Kepler-155 c	52.66153	$1.87^{+0.11}_{-0.12}$	$1.28^{+0.26}_{-0.25}$	3984^{+71}_{-86}	$0.54^{+0.03}_{-0.04}$	30.6	1.000
463.01[c]	8845205	Kepler-560 b	18.47763	$1.55^{+0.32}_{-0.29}$	$1.21^{+0.72}_{-0.47}$	3395^{+74}_{-67}	$0.28^{+0.06}_{-0.05}$	78.0	0.001
494.01	3966801	Kepler-577 b	25.69581	$1.70^{+0.21}_{-0.33}$	$2.30^{+1.17}_{-1.10}$	3787^{+163}_{-204}	$0.48^{+0.06}_{-0.09}$	35.9	1.000
571.05[a]	8120608	Kepler-186 f	129.94410	$1.18^{+0.11}_{-0.14}$	$0.23^{+0.07}_{-0.06}$	3751^{+75}_{-84}	$0.44^{+0.04}_{-0.05}$	7.7	0.677
701.03	9002278	Kepler-62 e	122.38740	$1.72^{+0.10}_{-0.07}$	$1.24^{+0.27}_{-0.19}$	4926^{+98}_{-98}	$0.66^{+0.04}_{-0.03}$	35.9	0.994
701.04[d]	9002278	Kepler-62 f	267.29100	$1.43^{+0.08}_{-0.06}$	$0.44^{+0.09}_{-0.07}$	4926^{+98}_{-98}	$0.66^{+0.04}_{-0.03}$	14.3	0.000
812.03	4139816	Kepler-235 e	46.18420	$1.83^{+0.12}_{-0.15}$	$1.32^{+0.29}_{-0.30}$	3950^{+70}_{-86}	$0.49^{+0.03}_{-0.04}$	18.0	1.000
854.01	6435936	Kepler-705 b	56.05608	$1.94^{+0.12}_{-0.22}$	$0.69^{+0.15}_{-0.19}$	3593^{+71}_{-86}	$0.49^{+0.03}_{-0.06}$	19.3	0.996
947.01	9710326	Kepler-737 b	28.59914	$1.83^{+0.16}_{-0.21}$	$1.87^{+0.52}_{-0.53}$	3755^{+75}_{-84}	$0.46^{+0.04}_{-0.05}$	45.7	1.000
1078.03	10166274	Kepler-267 d	28.46465	$1.87^{+0.14}_{-0.22}$	$1.95^{+0.49}_{-0.55}$	3789^{+75}_{-82}	$0.46^{+0.04}_{-0.05}$	22.2	0.992
1298.02[d]	10604335	Kepler-283 c	92.74958	$1.87^{+0.08}_{-0.10}$	$0.78^{+0.15}_{-0.14}$	4141^{+83}_{-91}	$0.58^{+0.03}_{-0.03}$	10.7	0.000
1404.02	8874090	···	18.90609	$0.87^{+0.16}_{-0.21}$	$3.03^{+2.29}_{-1.67}$	3751^{+219}_{-219}	$0.45^{+0.08}_{-0.11}$	10.1	0.955
1422.02[b]	11497958	Kepler-296 d	19.85029	$1.52^{+0.19}_{-0.23}$	$1.83^{+0.68}_{-0.62}$	3526^{+71}_{-78}	$0.38^{+0.05}_{-0.06}$	25.1	1.000
1422.04	11497958	Kepler-296 f	63.33627	$1.18^{+0.15}_{-0.18}$	$0.39^{+0.15}_{-0.13}$	3526^{+71}_{-78}	$0.38^{+0.05}_{-0.06}$	9.1	0.927
1422.05	11497958	Kepler-296 e	34.14211	$1.06^{+0.13}_{-0.16}$	$0.89^{+0.33}_{-0.30}$	3526^{+71}_{-78}	$0.38^{+0.05}_{-0.06}$	10.5	0.984
1596.02	10027323	Kepler-309 c	105.35823	$1.87^{+0.13}_{-0.17}$	$0.41^{+0.09}_{-0.10}$	3883^{+69}_{-93}	$0.50^{+0.04}_{-0.04}$	16.5	0.738
2162.02	9205938	···	199.66876	$1.45^{+0.18}_{-0.18}$	$2.06^{+0.76}_{-0.59}$	5678^{+113}_{-102}	$0.92^{+0.12}_{-0.12}$	11.1	0.920
2184.02[e]	12885212	···	95.90640	$2.17^{+0.07}_{-0.12}$	$1.63^{+0.20}_{-0.29}$	4620^{+73}_{-82}	$0.74^{+0.02}_{-0.04}$	8.92	0.638
2418.01	10027247	Kepler-1229 b	86.82952	$1.68^{+0.12}_{-0.21}$	$0.35^{+0.08}_{-0.11}$	3576^{+71}_{-85}	$0.46^{+0.03}_{-0.06}$	11.7	0.937
2626.01	11768142	···	38.09707	$1.58^{+0.20}_{-0.21}$	$0.81^{+0.30}_{-0.25}$	3554^{+71}_{-80}	$0.40^{+0.05}_{-0.05}$	14.6	0.999
2650.01	8890150	Kepler-395 c	34.98978	$1.14^{+0.07}_{-0.10}$	$1.71^{+0.35}_{-0.42}$	3765^{+75}_{-83}	$0.52^{+0.03}_{-0.05}$	10.1	0.985
2719.02	5184911	···	106.25976	$1.50^{+0.10}_{-0.16}$	$1.99^{+0.53}_{-0.58}$	4827^{+129}_{-144}	$0.82^{+0.06}_{-0.09}$	10.0	0.990
3010.01	3642335	Kepler-1410 b	60.86610	$1.39^{+0.07}_{-0.10}$	$0.84^{+0.17}_{-0.16}$	3808^{+69}_{-76}	$0.52^{+0.03}_{-0.04}$	12.7	0.996
3034.01	2973386	···	31.02092	$1.66^{+0.12}_{-0.17}$	$1.70^{+0.40}_{-0.45}$	3720^{+73}_{-81}	$0.48^{+0.03}_{-0.05}$	11.9	1.000
3138.01[b]	6444896	Kepler-1649 b	8.68909	$0.49^{+0.00}_{-0.00}$	$0.47^{+0.00}_{-0.00}$	2703^{+0}_{-0}	$0.12^{+0.00}_{-0.00}$	12.0	1.000
3282.01	12066569	Kepler-1455 b	49.27684	$1.75^{+0.09}_{-0.13}$	$1.28^{+0.26}_{-0.26}$	3899^{+78}_{-78}	$0.53^{+0.03}_{-0.04}$	14.7	0.996
3284.01	6497146	Kepler-438 b	35.23319	$0.97^{+0.06}_{-0.07}$	$1.62^{+0.37}_{-0.34}$	3749^{+75}_{-84}	$0.52^{+0.03}_{-0.04}$	11.9	1.000
3497.01	8424002	Kepler-1512 b	20.35972	$0.80^{+0.12}_{-0.16}$	$1.38^{+0.58}_{-0.58}$	3419^{+67}_{-76}	$0.34^{+0.05}_{-0.07}$	19.6	1.000
4005.01[a]	8142787	Kepler-439 b	178.13960	$2.25^{+0.22}_{-0.16}$	$1.70^{+0.47}_{-0.31}$	5431^{+81}_{-81}	$0.88^{+0.09}_{-0.06}$	17.8	0.997
4036.01	11415243	Kepler-1544 b	168.81133	$1.69^{+0.10}_{-0.06}$	$0.80^{+0.17}_{-0.12}$	4798^{+95}_{-95}	$0.71^{+0.04}_{-0.03}$	14.8	0.965
4087.01	6106282	Kepler-440 b	101.11141	$1.61^{+0.10}_{-0.08}$	$0.65^{+0.14}_{-0.11}$	4133^{+74}_{-82}	$0.56^{+0.03}_{-0.03}$	15.7	1.000
4356.01[a]	8459663	Kepler-1593 b	174.51028	$1.74^{+0.14}_{-0.20}$	$0.28^{+0.09}_{-0.09}$	4367^{+124}_{-155}	$0.45^{+0.04}_{-0.05}$	11.0	0.976
4427.01	4172805	···	147.66173	$1.59^{+0.12}_{-0.14}$	$0.23^{+0.06}_{-0.05}$	3788^{+76}_{-84}	$0.49^{+0.04}_{-0.04}$	10.8	0.969
4460.01	9947389	···	284.72721	$2.02^{+0.30}_{-0.29}$	$1.41^{+0.55}_{-0.44}$	5497^{+82}_{-74}	$1.08^{+0.16}_{-0.16}$	10.7	0.972
4550.01	5977470	···	140.25194	$1.84^{+0.05}_{-0.12}$	$1.28^{+0.17}_{-0.24}$	4821^{+76}_{-86}	$0.79^{+0.02}_{-0.05}$	9.6	0.934
4622.01	11284772	Kepler-441 b	207.24820	$1.56^{+0.09}_{-0.06}$	$0.30^{+0.06}_{-0.05}$	4339^{+78}_{-87}	$0.55^{+0.03}_{-0.02}$	9.7	0.975
4742.01	4138008	Kepler-442 b	112.30530	$1.30^{+0.07}_{-0.05}$	$0.79^{+0.15}_{-0.11}$	4401^{+78}_{-78}	$0.59^{+0.03}_{-0.02}$	12.9	0.993
7016.01	8311864	Kepler-452 b	384.84300	$1.09^{+0.20}_{-0.10}$	$0.56^{+0.32}_{-0.15}$	5579^{+150}_{-150}	$0.80^{+0.15}_{-0.07}$	7.6	0.771
7223.01	9674320	···	317.06242	$1.59^{+0.27}_{-0.12}$	$0.54^{+0.29}_{-0.13}$	5366^{+160}_{-144}	$0.71^{+0.12}_{-0.05}$	10.3	0.947
7706.01	4762283	···	42.04952	$1.19^{+0.08}_{-0.16}$	$2.00^{+0.55}_{-0.68}$	4281^{+115}_{-140}	$0.48^{+0.03}_{-0.06}$	7.5	0.837
7711.01	4940203	···	302.77982	$1.31^{+0.34}_{-0.12}$	$0.87^{+0.66}_{-0.22}$	5734^{+154}_{-154}	$0.80^{+0.21}_{-0.07}$	8.5	0.987
7882.01	8364232	···	65.41518	$1.31^{+0.08}_{-0.12}$	$1.79^{+0.49}_{-0.47}$	4348^{+130}_{-130}	$0.65^{+0.04}_{-0.06}$	7.2	0.529
7894.01	8555967	···	347.97611	$1.62^{+0.49}_{-0.15}$	$0.97^{+0.87}_{-0.27}$	5995^{+163}_{-181}	$0.88^{+0.27}_{-0.08}$	8.5	0.837
7923.01	9084569	···	395.13138	$0.97^{+0.12}_{-0.10}$	$0.44^{+0.20}_{-0.13}$	5060^{+192}_{-174}	$0.87^{+0.10}_{-0.09}$	10.0	0.750
7954.01	9650762	···	372.15035	$1.74^{+0.46}_{-0.14}$	$0.69^{+0.52}_{-0.18}$	5769^{+155}_{-172}	$0.81^{+0.21}_{-0.07}$	8.9	0.839
8000.01	10331279	···	225.48805	$1.70^{+0.43}_{-0.14}$	$1.20^{+0.90}_{-0.30}$	5663^{+169}_{-152}	$0.78^{+0.19}_{-0.07}$	8.7	0.975
8012.01	10452252	···	34.57372	$0.42^{+0.17}_{-0.12}$	$0.37^{+0.47}_{-0.19}$	3374^{+112}_{-82}	$0.22^{+0.09}_{-0.06}$	7.7	0.989
8174.01	8873873	···	295.06066	$0.64^{+0.07}_{-0.07}$	$0.70^{+0.28}_{-0.21}$	5332^{+160}_{-144}	$0.76^{+0.09}_{-0.09}$	7.4	0.665

[a] Confirmed planet properties from NASA Exoplanet Archive on May 31, 2017 place object within HZ.

[b] Confirmed planet properties from NASA Exoplanet Archive on May 31, 2017 place object exterior to the HZ.

[c] Confirmed planet with vetting score less than 0.5.

[d] Confirmed planet dispositioned as False Positive in DR25.

[e] The erratum to Mathur et al. (2017) reduces planet size, now placing the object in the eta-Earth sample.

radius distribution of exoplanets for planets in orbital periods of less than 100 d. The authors argue that the gap is the result of two (possibly overlapping) population distributions: the rocky terrestrials and the mini-Neptune size planets characterized by their volatile-rich envelopes. Within this framework, the center of the gap marks a probabilistic boundary between having a higher likelihood of a terrestrial composition versus a higher likelihood of a volatile-rich envelope. However, this boundary was identified using planets in orbital periods of less than 100 days and it may not exist for planets in longer period orbits. Also, it is not entirely clear that planets on the small side of this gap are all terrestrial. Rogers (2015) examined small planets with density measurements with periods less than ≈50 d and showed that less than half of planets with a radii of 1.62 $R_\oplus$ have densities consistent with a body primarily composed of iron and silicates. For our purposes of highlighting the smallest planets in this catalog, we chose to be inclusive and set the threshold at 1.8 $R_\oplus$.

To summarize, Table 7 lists those candidates with scores greater than 0.5 and whose error bars indicate that they could be smaller than 1.8 $R_\oplus$ and lie in the habitable zone. The table also includes KOI 2184.02 because the erratum to Mathur et al. (2017, see §2.5 of this paper) reduces the stellar and planet radii so that the PC now lies in our sample. Note, the same erratum also reduces the planet radii of KOI 4460.01 and KOI 4550.01 to 2.0 $R_\oplus$ and 1.65 $R_\oplus$ respectively. The values reported in Table 7 are identical to those in the KOI table at the NASA Exoplanet Archive and do not include the values reported in the erratum to Mathur et al. (2017). Also, in order to make Table 7 complete we include any *Kepler* terrestrial-size confirmed planet that falls in the habitable zone of its star according to the confirmed planet table at the Exoplanet Archive (downloaded on 2017-05-15). The objects are included, and denoted with footnotes, even if the DR25 catalog dispositions them as FPs, or if the DR25 planetary parameters place them outside the habitable zone. However, note that statistical inferences like occurrence rates should be based on a uniform sample drawn exclusively from the DR25 catalog and its self-consistent completeness and reliability measurements (see §8).

We plot the eta-Earth sample candidates in Figure 14, using only the information in the DR25 KOI catalog. Notice that this final search of the *Kepler* data not only identified previously discovered candidates around the M dwarf stars, it also yielded a handful of highly reliable candidates around the GK dwarf stars. These GK dwarf candidates have fewer transits and shallower depths, making them much more difficult to find. Despite their lower signal-to-noise, because we provide a measure of the reliability against false alarms (along with the completeness), these candidates are available to further study the occurrence rates of small planets in the habitable zone of GK dwarf stars.

7.5.2. *Notes on the Eta-Earth Sample*

Forty-seven candidates have a score greater than 0.5 and fall in this eta-Earth sample; 10 of these are new to this catalog (KOI numbers greater than 7621.01 and KOI 238.03). A manual review of the 10 new high-score candidates indicates that they are all low signal-to-noise with very few transits, and show no obvious reason to be called false positives. However, our reliability measurements indicate that ≈20% of these targets are not caused by a transiting/eclipsing system. As an example, the candidate most similar to the size and temperature of the Earth is KOI 7711.01 (KIC 004940203), with four transits that all cleanly pass the individual transit metrics. It orbits a 5734 K star, has an insolation flux slightly less than that of Earth, and is about 30% larger according to its DR25 catalog properties. Plots showing visualizations of the transit data and its quality are available at the Exoplanet Archive for this object[18] and for all of the obsTCEs, injTCEs, scrTCEs, and invTCEs.

Several confirmed planets fall in our eta-Earth sample. Kepler-186f (KOI 571.05), Kepler-439b (KOI 4005.01) and Kepler-1593b (KOI 4356.01) move into the habitable zone according to the confirmed planet properties. They are included in Table 7 with a footnote indicating they would not otherwise be listed. Kepler-296d (KOI 1422.02) and Kepler-1649b (KOI 3138.01), on the other hand, move outside the HZ according to the updated properties and are noted accordingly. Note, the default properties in the confirmed planets table at the Exoplanet Archive are selected for completeness and precision. Additional values may be available from other references that represent the best, current state of our knowledge.

Kepler-560b (KOI 463.01) is a confirmed planet that is a PC in the DR25 catalog, but failed the score cut; it is included for awareness and annotated accordingly. The low score is caused by the Centroid Robovetter (§A.5.1) detecting a possible offset from the star's cataloged position, likely due to the star's high proper motion (Mann et al. 2017).

Two confirmed planets dispositioned as FPs in the DR25 catalog are included in Table 7: Kepler-62f (KOI 701.04) and Kepler-283c (KOI 1298.02). Kepler-62f has only 4 transit events in the time series. The transit observed during Quarter 9 is on the edge of a gap and narrowly fails Rubble. The transit observed during Quarter 12 is flagged by the Skye metric. Taken together, this leaves fewer than three unequivocal transits, the minimum required for the PC disposition.

Kepler-283c (KOI 1298.02) fails the shape metric. Its phase-folded transit appears v-shaped when TTVs are

[18] https://exoplanetarchive.ipac.caltech.edu/data/KeplerData/\penalty\z@004/004940/004940203/tcert/kplr004940203_q1_q17_dr25_obs_tcert.pdf

not included in the modeling. We note that vetting metrics employed by the DR25 Robovetter were computed without consideration of transit timing variations, whereas the transit fits used in the KOI table, described in §6.3, includes the timing variations as measured by Rowe et al. (2015a).

7.6. *Caveats*

When selecting candidates from the KOI catalog for further study, as we did for the eta-Earth sample (§7.5), it is important to remember a few caveats. First, even with a high cut on disposition score, the reliability against false alarms is not 100%. Some candidates may still be caused by false alarms, especially those around the larger, hotter stars. Also, this reliability number does not include the astrophysical reliability. Many of our tools to detect astrophysical false positives do not work for long-period, low MES candidates. For example, it is nearly impossible to detect the centroid offset created from a background eclipsing binary and secondary eclipses are not deep enough to detect for these stars.

Second, the measured radius and semi-major axis of each planet depends on the stellar catalog. As discussed in §2.5 and Mathur et al. (2017), the stellar radii and masses are only known to a certain precision and the quality of the data used to derive these stellar properties varies between targets. These unknowns are reflected in the 1-sigma error bars shown in Figure 14 and listed in the KOI table. The uncertainty in the stellar information limits our knowledge of these planets. As an example, for Kepler-452 (KIC 8311864), the DR25 stellar catalog lists a temperature of 5579±150 K and stellar radius of $0.798^{+0.150}_{-0.075}\,R_\odot$, while the values in the confirmation paper (Jenkins et al. 2015) after extensive follow-up are 5757±85 K for the effective temperature and $1.11^{+0.15}_{-0.09}$ for the stellar radius. As a result, the planet Kepler-452b is given as $1.6\pm0.2\,R_\oplus$ in Jenkins et al. (2015) and $1.09^{+0.2}_{-0.1}\,R_\oplus$ in the DR25 catalog. The radii and stellar temperature differ by less than 2-sigma, but those differences change the interpretation of the planet from a super-Earth in the middle of the habitable zone of an early G dwarf host to an Earth-size planet receiving about half the amount of flux from a late K star. As follow-up observations of each candidate star is obtained and errors on the stellar parameters decrease, we expect this population to change in significant ways.

Third, high-resolution imaging has proven crucial for identifying light from background and bound stars which add flux to the *Kepler* photometric time series (Furlan et al. 2017). When this occurs, unaccounted for extra light dilutes the transit, causing the radii to be significantly underestimated (Ciardi et al. 2015; Furlan & Howell 2017). As a result, we fully expect that once follow-up observations are obtained for these stars, several of the PCs in this catalog, including those listed in the eta-Earth sample, will be found to have radii larger than reported in this catalog.

8. USING THE DR25 CATALOG FOR OCCURRENCE RATE CALCULATIONS

The DR25 candidate catalog was designed with the goal of providing a well characterized sample of planetary candidates for use in occurrence rate calculations. For those smallest planets at the longest periods, our vetting is especially prone to miss transits and confuse other signals as transits, and this must be accounted for when doing occurrence rates. However, the completeness and reliability presented in this paper are simply the last two pieces of a much larger puzzle that must be assembled in order to perform occurrence rates with this catalog. In this section we endeavor to make users aware of other issues and biases, as well as all the products available to help interpret this KOI catalog, all of which are hosted at the NASA exoplanet archive.

8.1. *Pipeline Detection Efficiency*

Any measure of the catalog completeness must include the completeness of the Robovetter and the *Kepler* Pipeline. The Pipeline's detection efficiency has been explored in two ways: using pixel-level transit injection and using flux-level transit injection. In the former, a simulated transiting planet signal is injected into the calibrated pixels of each Kepler target, which are then processed through the pipeline. This experiment provides an estimate of the average detection efficiency over all the stars that were searched. A full description of the signals that were injected and recovered can be found in Christiansen (2017). The pixel-level measurements have the advantage of following transit signals through all the processing steps of the *Kepler* Pipeline, and the recovered signals can be further classified with the Robovetter, as demonstrated in §7.3. Figure 15 shows the average pipeline detection efficiency for a sample of FGK stars: the left panel shows the pipeline detection efficiency, and the right panel shows the combined Pipeline and Robovetter detection efficiency, calculated by taking the injections that were successfully recovered by the pipeline and processing them through the Robovetter. A gamma cumulative distribution function is fit to both (see equation 1 of Christiansen et al. 2016). Notice that the detection efficiency decreases by 5–10 percentage points (of the entire set that were injected) for all MES, as expected given the results shown in Figure 9.

Since the pixel-level transit injection includes only one injection per target, it does not examine potential variations in the pipeline completeness for individual targets due to differences in stellar properties or astrophysical variability. To probe these variations, a small number of individual stars had a large number of transiting signals (either several thousand or several hundred thousand, depending on the analysis) injected into the detrended photometry, which was processed only through the transit-search portion of the TPS module. The flux-level injections revealed that there are significant target-to-target variations in the detection efficiency. The flux-

level injections and the resulting detection efficiency is available for the sample of stars that were part of this study. For more information on the flux-level injection study see Burke & Catanzarite (2017c). All products associated with the flux-level and pixel-level injections can be found at the NASA Exoplanet Archive.[19]

8.2. *Astrophysical Reliability*

We have described the reliability of the DR25 candidates with regard to the possibility that the observed events are actually caused by stellar or instrumental noise. See §7.3 for how this reliability varies with various measured parameters. However, even if the observed signal is not noise, other astrophysical events can mimic a transit. Some of these other astrophysical events are removed by carefully vetting the KOI with *Kepler* data alone. Specifically, the Robovetter looks for significant secondary eclipses to rule out eclipsing binaries, and for a significant offset in the location of the in- and out-of-transit centroids to rule-out background eclipsing binaries. Morton et al. (2016) developed the vespa tool which considers the likelihood that a transit event is caused by various astrophysical events, including a planet. The False Positive Probabilities (FPP) table[20] provides the results of applying this tool to the KOIs in the DR25 catalog. It provides a probability that the observed signal is one of the known types of astrophysical false positives. The FPP table results are only reliable for high signal-to-noise (MES$\gtrapprox$10) candidates with no evidence that the transit occurs on a background source. For more information on this table see the associated documentation at the NASA Exoplanet Archive.

To robustly determine whether a KOI's signal originates from the target star, see the Astrophysical Positional Probabilities Table[21]. Using a more complete catalog of stars than the original Kepler Input Catalog (Brown et al. 2011), Bryson & Morton (2017) calculates the probability that the observed transit-like signal originates from the target star. Note, these positional probabilities are computed independent of the results from the Centroid Robovetter, and are not used by the Robovetter.

To help understand the astrophysical reliability of the DR25 KOIs as a population, we have provided data to measure how well the Robovetter removes certain types of FPs. As part of the pixel-level transit injection efforts, we injected signals that mimic eclipsing binaries and background eclipsing binaries. Those that were recovered by the *Kepler* Pipeline can be used to measure the effectiveness of the Robovetter at removing this type of FP. A full description of these injections and an analysis of the Robovetter's effectiveness in detecting these signals can be found in Coughlin (2017b).

8.3. *Imperfect Stellar Information*

For those doing occurrence rates, another issue to consider is whether the measured size of the planet is correct. As discussed in §2.5, the stellar catalog (i.e., radii and temperatures) provided by Mathur et al. (2017) typically has errors of 27 percent for the stellar radii. Results from *Gaia* (Gaia Collaboration et al. 2016a,b) are expected to fix many of the shortcomings of this catalog. Also, the dilution from an unaccounted for bound or line-of-sight binary (Ciardi et al. 2015; Furlan et al. 2017), can cause planet radii to be larger than what is reported in the DR25 catalog. For occurrence rate calculations this dilution also has implications for the stars that have no observed planets because it means the search did not extend to planet radii that are as small as the stellar catalog indicates. For this reason, any correction to the occurrence rates that might be applied needs to consider the effect on all searched stars, not just the planet hosts.

9. CONCLUSIONS

The DR25 KOI catalog has been characterized so that it can serve as the basis for occurrence rate studies of exoplanets with periods as long as 500 days. The detection efficiency of the entire search (Burke & Catanzarite 2017a; Christiansen 2017) and of the Robovetter vetting process (Coughlin 2017b) has been calculated by injecting planetary transits into the data and determining which types of planets are found and which are missed. For this DR25 KOI catalog, the vetting completeness has been balanced against the catalog reliability, i.e., how often false alarms are mistakenly classified as PCs. This is the first *Kepler* exoplanet catalog to be characterized in this way, enabling occurrence rate measurements at the detection limit of the mission. As a result, accurate measurements of the frequency of terrestrial-size planets at orbital periods of hundreds of days is possible.

The measurement of the reliability using the inverted and scrambled light curves is new to this KOI catalog. We measure how often noise is labeled as a planet candidate and combine that information with the number of false alarms coming from the *Kepler* Pipeline. Some pure noise signals so closely mimic transiting signals that it is nearly impossible to remove them all. Because of this, it is absolutely imperative that those using this candidate catalog for occurrence rates consider this source of noise. For periods longer than ≈200 days and radii less than $\approx 4\,R_\oplus$, these noise events are often labelled as PC and thus the reliability of the catalog is near 50%. Astrophysical reliability is another concern that must be accounted for independently. However, even once it is shown that another astrophysical scenario is unlikely

[19] https://exoplanetarchive.ipac.caltech.edu/docs/KeplerSimulated.html

[20] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koifpp

[21] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koiapp

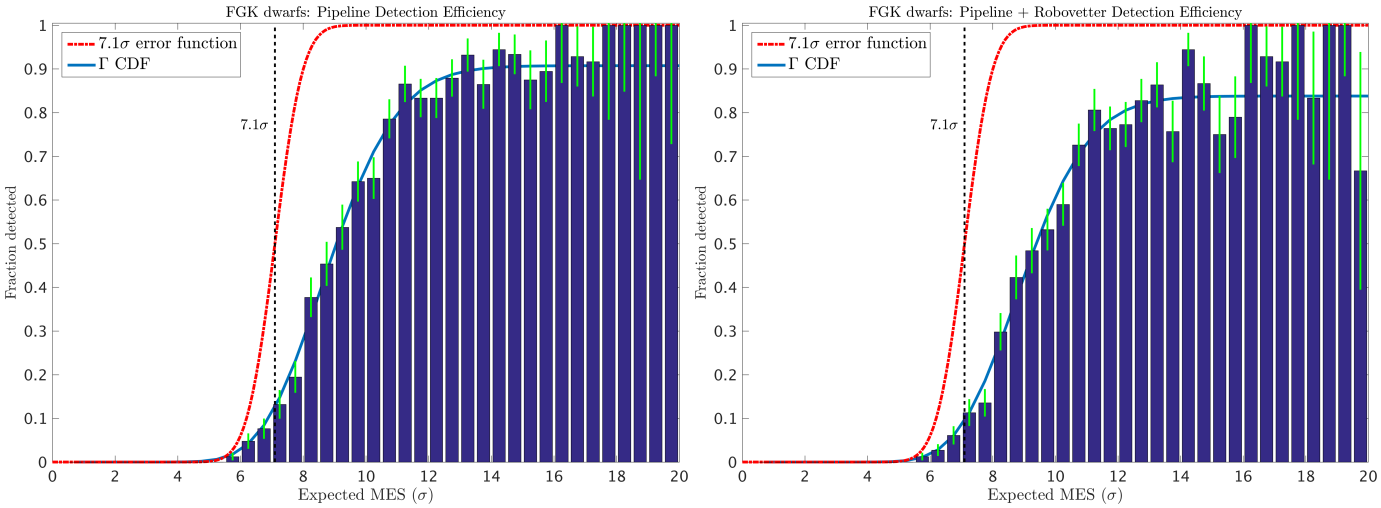

Figure 15. Left: The average detection efficiency of the *Kepler* Pipeline for a sample of FGK stars, as measured by the pixel-level transit injection experiment and described by Christiansen (2017). The solid blue line is a best-fit Γ cumulative distribution function (see Equation 1 of Christiansen et al. 2016); the red dashed line shows the hypothetical performance for a perfect detector in TPS. Right: The average detection efficiency of the *Kepler* Pipeline and the Robovetter, where the injections successfully recovered by the Pipeline are then subsequently evaluated as PCs by the Robovetter.

(as was done for the DR24 KOIs in Morton et al. 2016), the PCs in this catalog cannot be validated without first showing that the candidates have a sufficiently high false alarm reliability.

We have shown several ways to identify to identify high reliability or high completeness samples. Reliability is a strong function of the MES and the number of observed transits. Also, the FGK dwarf stars are known to be quieter than giant stars and in general the true transits can be more easily separated from the false alarms. We also provide the disposition score, a measure of how robustly a candidate has passed the Robovetter; this can be used to easily find the most reliable candidates. Those doing follow-up observations of KOIs may also use this disposition score to identify the candidates that will optimize ground-based follow-up observations.

This search of the *Kepler* data yielded 219 new PCs. Among those new candidates are two new candidates in multi-planet systems (KOI-82.06 and KOI-2926.05). Also, the catalog contains ten new high-reliability, super-Earth size, habitable zone candidates. Some of the most scrutinized signals in the DR25 KOI catalog will likely be those fifty small, temperate PCs in the eta-Earth sample defined in §7.5. These signals, along with their well characterized completeness and reliability, can be used to make an almost direct measurement on the occurrence rate of planets with the size and insolation flux as Earth, especially around GK dwarf stars. While this catalog is an important step forward in measuring this number, it is important to remember a few potential biases inherent to this catalog. Namely, errors in the stellar parameters result in significant errors on the planetary sizes and orbital distances, and unaccounted for background stars make planet radii appear smaller than reality and impact the detection limit of the search for all stars. Also, the Robovetter is not perfect — completeness of the vetting procedures and the reliability of these signals (both astrophysical and false alarm) must be considered in any calculation.

Ultimately, characterizing this catalog was made possible because of the Robovetter (§3) and the innovative metrics it uses to vet each TCE. It has improved the uniformity and accuracy of the vetting process and has allowed the entire process to be tested with known transits and known false positives. As a result, the Robovetter could be run many times, each time improving the vetting by changing thresholds or introducing new metrics. We adapted our vetting process as we learned about the data set, ensuring the highest reliability and completeness achievable in the time allowed. The Robovetter metrics and logic may prove useful for future transit missions that will find an unprecedented abundance of signals that will require rapid candidate identification for ground-based follow-up, e.g., K2 (Howell et al. 2014), TESS (Ricker et al. 2015), and PLATO (Rauer et al. 2016).

This paper includes data collected by the *Kepler* mission. The Kepler Mission was a PI-led Discovery Class Mission funded by the NASA Science Mission directorate. The authors acknowledge the efforts of the *Kepler* Mission team for generating the many data products used to create the KOI catalog. These products were generated by the *Kepler* Mission science pipeline through the efforts of the Kepler Science Operations Center and Science Office. The Kepler Mission is led by the project office at NASA Ames Research Center. Ball

Aerospace built the Kepler photometer and spacecraft which is operated by the mission operations center at LASP. We acknowledge the Kepler Education and Outreach team, including Alan Gould, Edna DeVore and Michele Johnson, for their efforts in making the results of this paper accessible to the public. We thank the many scientists who have contributed to the *Kepler* Mission over the years, including R. Gilliland, E. Furlan, J. Orosz and K. Colón. We thank the managers and engineers who worked on *Kepler* over the years, without whom we would not have had a successful *Kepler* Mission. This research has made use of NASA's Astrophysics Data System. We thank GNU parallel for enabling rapid running of the Robovetter input metrics (Tange 2011). We thank P.P. Mullally for inspiring the names of certain algorithms. Thank you to Turbo-King et al. (2017) for a spirited discussion. Some of the data products used in this paper are archived at the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. Some of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts. J.F.R acknowledges support from NASA grant NNX14AB82G issued through the Kepler Participating Scientist Program. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. This research was enabled, in part, by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). D.H. and S.M. acknowledge support by the National Aeronautics and Space Administration under Grant NNX14AB92G issued through the Kepler Participating Scientist Program. J.L.C. is supported by NASA under award No. GRNASM99G000001. J.S. is supported by the NASA *Kepler* Participating Scientist Program NNX16AK32G. W.F.W. gratefully acknowledges support from NASA via the *Kepler* Participating Scientist Program grant NNX14AB91G. V.S.A. acknowledges support from VILLUM FONDEN (research grant 10118). Funding for the Stellar Astrophysics Centre is provided by The Danish National Research Foundation (Grant DNRF106). The research was supported by the ASTERISK project (ASTERoseismic Investigations with SONG and Kepler) funded by the European Research Council (Grant agreement no.: 267864).

REFERENCES

Aigrain, S., Llama, J., Ceillier, T., et al. 2015, MNRAS, 450, 3211

Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, PASP, 125, 989

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2014, ArXiv e-prints, arXiv:1403.6015

Barclay, T., Rowe, J. F., Lissauer, J. J., et al. 2013, Nature, 494, 452

Barge, P., Baglin, A., Auvergne, M., et al. 2008, A&A, 482, L17

Batalha, N. M., Borucki, W. J., Bryson, S. T., et al. 2011, ApJ, 729, 27

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, ApJS, 204, 24

Borucki, W. J. 2016, Reports on Progress in Physics, 79, 036901

Borucki, W. J., Koch, D., Jenkins, J., et al. 2009, Science, 325, 709

Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, AJ, 142, 112

Bryson, S. T., & Morton, T. D. 2017, Planet Reliability Metrics: Astrophysical Positional Probabilities for Data Release 25 (KSCI-19108-001)

Bryson, S. T., Jenkins, J. M., Klaus, T. C., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al. 2013, PASP, 125, 889

Bryson, S. T., Abdul-Masih, M., Batalha, N., et al. 2017, The Kepler Certified False Positive Table (KSCI-19093-003)

Burke, C. J., & Catanzarite, J. 2017a, Planet Detection Metrics: Per-Target Detection Contours for Data Release 25 (KSCI-19111-002)

—. 2017b, Planet Detection Metrics: Per-Target Flux-Level Transit Injection Tests of TPS for Data Release 25 (KSCI-19109-002)

—. 2017c, Planet Detection Metrics: Window and One-Sigma Depth Functions for Data Release 25 (KSCI-19101-002)

Burke, C. J., & Seader, S. E. 2016, Window and One-Sigma Depth Functions for Data Release 24 (KSCI-19085-002)

Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8

Byrd, R. H., Peihuang, L., Nocedal, J., & Zhu, C. 1995, SIAM J. Sci. Comput., 16
Cacciari, C. 2009, Mem. Soc. Astron. Italiana, 80, 97
Chiavassa, A., Caldas, A., Selsis, F., et al. 2017, A&A, 597, A94
Christiansen, J. L. 2015, Planet Detection Metrics: Pipeline Detection Efficiency (KSCI-19094-001)
—. 2017, Planet Detection Metrics: Pixel-Level Transit Injection Tests of Pipeline Detection Efficiency for Data Release 25 (KSCI-19110-001)
Christiansen, J. L., Ballard, S., Charbonneau, D., et al. 2010, ApJ, 710, 97
Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, PASP, 124, 1279
—. 2013a, Kepler Data Characteristics Handbook (KSCI-19040-004), http://archive.stsci.edu/kepler/manuals/Data_Characteristics.pdf
Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013b, ApJS, 207, 35
—. 2015, ApJ, 810, 95
—. 2016, ApJ, 828, 99
Ciardi, D. R., Beichman, C. A., Horch, E. P., & Howell, S. B. 2015, ApJ, 805, 16
Claret, A. 2000, A&A, 359, 289
Claret, A., & Bloemen, S. 2011, A&A, 529, A75
Coughlin, J. L. 2014, Description of the TCERT Vetting Products for the Q1-Q16 Catalog Using SOC 9.1 (KSCI-19103-001)
—. 2017a, Description of the TCERT Vetting Reports for Data Release 25 (KSCI-19105-001)
—. 2017b, Planet Detection Metrics: Robovetter Completeness and Effectiveness for Data Release 25 (KSCI-19114-001)
Coughlin, J. L., & López-Morales, M. 2012, AJ, 143, 39
Coughlin, J. L., Thompson, S. E., Bryson, S. T., et al. 2014, AJ, 147, 119
Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, ApJS, 224, 12
Devor, J., Charbonneau, D., O'Donovan, F. T., Mandushev, G., & Torres, G. 2008, AJ, 135, 850
Dotter, A., Chaboyer, B., Jevremović, D., et al. 2008, ApJS, 178, 89
Doyle, L. R., Carter, J. A., Fabrycky, D. C., et al. 2011, Science, 333, 1602
Dressing, C. D., & Charbonneau, D. 2013, ApJ, 767, 95
—. 2015, ApJ, 807, 45
Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2014, ApJ, 790, 146
Fogtmann-Schulz, A., Hinrup, B., Van Eylen, V., et al. 2014, ApJ, 781, 67
Ford, E. B. 2005, AJ, 129, 1706
Foreman-Mackey, D., Morton, T. D., Hogg, D. W., Agol, E., & Schölkopf, B. 2016, AJ, 152, 206
Fuller, J., Hambleton, K., Shporer, A., Isaacson, H., & Thompson, S. 2017, ArXiv e-prints, arXiv:1706.05053
Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, ArXiv e-prints, arXiv:1703.10375
Furlan, E., & Howell, S. B. 2017, The Astronomical Journal, 154, 66
Furlan, E., Ciardi, D. R., Everett, M. E., et al. 2017, AJ, 153, 71
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016a, A&A, 595, A2
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016b, A&A, 595, A1
Garcia, D. 2010, Computational Statistics & Data Analysis, 54, 1167
García, R. A., Ceillier, T., Salabert, D., et al. 2014, A&A, 572, A34
Gilliland, R. L., Cartier, K. M. S., Adams, E. R., et al. 2015, AJ, 149, 24
Gilliland, R. L., Chaplin, W. J., Dunham, E. W., et al. 2011, ApJS, 197, 6
Hambleton, K., Fuller, J., Thompson, S., et al. 2017, ArXiv e-prints, arXiv:1706.05051
Hampel, F. R. 1974, Journal of the American Statistical Association, 69, 383
He, X., & Niyogi, P. 2004, Advances in Neural Inoformation Processing Systems, 16, 37
Hoffman, K. L., & Rowe, J. F. 2017, Uniform Modeling of KOIs: MCMC Notes for Data Release 25 (KSCI-19113-001)
Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15
Howell, S. B., Ciardi, D. R., Giampapa, M. S., et al. 2016, AJ, 151, 43
Howell, S. B., Sobeck, C., Haas, M., et al. 2014, PASP, 126, 398
Huber, D., Silva Aguirre, V., Matthews, J. M., et al. 2014, ApJS, 211, 2
Jenkins, J. M. 2002, ApJ, 575, 493
—. 2017a, Kepler Data Processing Handbook (KSCI-19081-002)
—. 2017b, Kepler Data Processing Handbook (KSCI-19081-002)
Jenkins, J. M., Caldwell, D. A., & Borucki, W. J. 2002, ApJ, 564, 495
Jenkins, J. M., Twicken, J. D., Batalha, N. M., et al. 2015, AJ, 150, 56

Johnson, J. A., Petigura, E. A., Fulton, B. J., et al. 2017, ArXiv e-prints, arXiv:1703.10402
Kirk, B., Conroy, K., Prša, A., et al. 2016, AJ, 151, 68
Koch, D. G., Borucki, W. J., Rowe, J. F., et al. 2010, ApJL, 713, L131
Kopparapu, R. K., Ramirez, R., Kasting, J. F., et al. 2013, ApJ, 765, 131
Kreiner, J. M. 2004, AcA, 54, 207
Kruse, E., & Agol, E. 2014, Science, 344, 275
Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al. 2011a, Nature, 470, 53
Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., et al. 2011b, ApJS, 197, 8
Lissauer, J. J., Marcy, G. W., Bryson, S. T., et al. 2014, ApJ, 784, 44
Lopez, E. D., & Fortney, J. J. 2013, ApJ, 776, 2
Lundkvist, M. S., Kjeldsen, H., Albrecht, S., et al. 2016, Nature Communications, 7, 11201
Mandel, K., & Agol, E. 2002, ApJL, 580, L171
Mann, A. W., Dupuy, T., Muirhead, P. S., et al. 2017, AJ, 153, 267
Marcy, G. W., Isaacson, H., Howard, A. W., et al. 2014, ApJS, 210, 20
Mathur, S., Huber, D., Batalha, N. M., et al. 2017, ApJS, 229, 30
Mayor, M., & Queloz, D. 1995, Nature, 378, 355
Mazeh, T., Nachmani, G., Sokol, G., Faigler, S., & Zucker, S. 2012, A&A, 541, A56
McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24
Meibom, S., Mathieu, R. D., Stassun, K. G., Liebesny, P., & Saar, S. H. 2011, ApJ, 733, 115
Mignard, F. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 338, Astrometry in the Age of the Next Generation of Large Telescopes, ed. P. K. Seidelmann & A. K. B. Monet, 15–+
More, J., Garbow, B., & Hillstrom, K. 1980, Argoone National Laboratory Report ANL-80-74
Morton, T. D., Bryson, S. T., Coughlin, J. L., et al. 2016, ApJ, 822, 86
Mullally, F. 2017, Planet Detection Metrics: Automatic Detection of Background Objects Using the Centroid Robovetter (KSCI-19115-001)
Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2016, PASP, 128, 074502
—. 2015, ApJS, 217, 31
O'Donovan, F. T., Charbonneau, D., Mandushev, G., et al. 2006, ApJL, 651, L61
Orosz, J. A., Welsh, W. F., Carter, J. A., et al. 2012, Science, 337, 1511
Owen, J. E., & Wu, Y. 2013, ApJ, 775, 105
Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, Proceedings of the National Academy of Science, 110, 19273
Petigura, E. A., Howard, A. W., Marcy, G. W., et al. 2017a, AJ, 154, 107
—. 2017b, ArXiv e-prints, arXiv:1703.10400
Prša, A., Batalha, N., Slawson, R. W., et al. 2011, AJ, 141, 83
Quintana, E. V., Barclay, T., Raymond, S. N., et al. 2014, Science, 344, 277
Rappaport, S., Vanderburg, A., Jacobs, T., et al. 2017, ArXiv e-prints, arXiv:1708.06069
Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian Processes for Machine Learning (The MIT Press)
Rauer, H., Aerts, C., Cabrera, J., & PLATO Team. 2016, Astronomische Nachrichten, 337, 961
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003
Rogers, L. A. 2015, ApJ, 801, 41
Rowe, J. F., Bryson, S. T., Marcy, G. W., et al. 2014, ApJ, 784, 45
Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015a, ApJS, 217, 16
—. 2015b, ApJS, 217, 16
Ruppert, D. 2010, Statistics and Data Analysis for Financial Engineering (Springer Texts in Statistics), 1st edn. (Springer, Berlin)
Samus, N. N., Durlevich, O. V., & et al. 2009, VizieR Online Data Catalog, 1, 2025
Santerne, A., Moutou, C., Tsantaki, M., et al. 2016, A&A, 587, A64
Seader, S., Jenkins, J. M., Tenenbaum, P., et al. 2015, ApJS, 217, 18
Shporer, A. 2017, PASP, 129, 072001
Shporer, A., Jenkins, J. M., Rowe, J. F., et al. 2011, AJ, 142, 195
Shporer, A., Fuller, J., Isaacson, H., et al. 2016, ApJ, 829, 34
Slawson, R. W., Prša, A., Welsh, W. F., et al. 2011, AJ, 142, 160
Stark, C. C., Roberge, A., Mandell, A., et al. 2015, ApJ, 808, 149
Stumpe, M. C., Smith, J. C., Catanzarite, J. H., et al. 2014, PASP, 126, 100
Tange, O. 2011, ;login: The USENIX Magazine, 36, 42
Tenenbaum, P., Christiansen, J. L., Jenkins, J. M., et al. 2012, ApJS, 199, 24

Thompson, S. E., Fraquelli, D., van Cleve, J. E., & Caldwell, D. A. 2016a, Kepler Archive Manual (KDMC-10008-006)

Thompson, S. E., Mullally, F., Coughlin, J. L., et al. 2015, ApJ, 812, 46

Thompson, S. E., Everett, M., Mullally, F., et al. 2012, ApJ, 753, 86

Thompson, S. E., Caldwell, D. A., Jenkins, J. M., et al. 2016b, Kepler Data Release 25 Notes (KSCI-19065-002), http://archive.stsci.edu/kepler/release_notes/release_notes25/KSCI-19065-002DRN25.pdf

Torres, G., Kipping, D. M., Fressin, F., et al. 2015, ApJ, 800, 99

Turbo-King, M., Tang, B. R., Habeertable, Z., et al. 2017, ArXiv e-prints, arXiv:1703.10803

Twicken, J. D., Jenkins, J. M., Seader, S. E., et al. 2016, AJ, 152, 158

Van Cleve, J. E., & Caldwell, D. A. 2009, Kepler Instrument Handbook (KSCI-19033-001), http://archive.stsci.edu/kepler/manuals/KSCI-19033-001.pdf

—. 2016, Kepler Instrument Handbook (KSCI-19033-0012, http://archive.stsci.edu/kepler/manuals/KSCI-19033-002.pdf

Van Cleve, J. E., Christiansen, J. L., Jenkins, J., & Caldwell, D. A. 2016a, Kepler Data Characteristics Handbook (KSCI-19040-005), http://archive.stsci.edu/kepler/manuals/Data_Characteristics.pdf

Van Cleve, J. E., Howell, S. B., Smith, J. C., et al. 2016b, PASP, 128, 075002

Welsh, W. F., Orosz, J. A., Aerts, C., et al. 2011, ApJS, 197, 4

Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Youdin, A. N. 2011, ApJ, 742, 38

Zimmerman, M., Thompson, S., Mullally, F., et al. 2017, ArXiv e-prints, arXiv:1706.05434

APPENDIX

A. ROBOVETTER METRIC DETAILS

In this appendix we describe, in detail, each of the Robovetter tests in the order in which they are performed by the Robovetter. See §3 for an overview of the logic used by the Robovetter.

A.1. *Two Robovetter Detrendings*

As mentioned in §1.2, for all of the Robovetter tests that require a phased light curve and model fit, we utilize two different detrendings and model fits (named ALT and DV). Both were also used by the DR24 Robovetter. Every test that is applied to the DV phased light curves is also applied to the ALT detrending, albeit with different thresholds for failure. Failing a test using either detrending results in the TCE being classified as an FP.

In the *Kepler* Pipeline, the DV module produces a harmonic-removed, median-detrended, phased flux light curve, along with a transit model fit (Jenkins 2017b; Wu et al. 2010). However, the harmonic removal software is known to suppress or distort short-period ($\lesssim$ 3 days) signals causing short-period eclipsing binaries with visible secondaries to appear as transiting planets with no visible secondaries (Christiansen et al. 2013b). It can also make variable stars with semi-coherent variability, such as star spots or pulsations, appear as transit-like signals. As an alternative, we implement the ALT detrending method that utilizes the pre-search data conditioned (PDC) time-series light curves and the non-parametric penalized least-squares detrending method of Garcia (2010) which includes only the out-of-transit points when computing the filter. This ALT detrending technique is effective at accurately detrending short-period eclipsing binaries and variable stars, i.e., preserving their astrophysical signal. These ALT detrended light curves are phased and fit with a simple trapezoidal transit model.

A.2. *The TCE is the Secondary of an Eclipsing Binary*

If a TCE under examination is not the first one in a system, the Robovetter checks if there exists a previous TCE with a similar period that was designated as an FP due to a stellar eclipse (see §A.4). (Note, TCEs for a given system are ordered from highest MES to lowest MES, and the Robovetter runs on them in this order.) To compute whether two TCEs have the same period within a given statistical threshold, we employ the period matching criteria of Coughlin et al. (2014, see equations 1-3), σ_P, where higher values of σ_P indicate more significant period matches. We re-state the equations here as:

$$\Delta P = \frac{P_A - P_B}{P_A} \tag{A1}$$

$$\Delta P' = \mathrm{abs}(\Delta P - \mathrm{rint}(\Delta P)) \tag{A2}$$

$$\sigma_P = \sqrt{2} \cdot \mathrm{erfcinv}(\Delta P') \tag{A3}$$

where P_A is the period of the shorter-period TCE, P_B is the period of the longer-period TCE, rint() rounds a number to the nearest integer, abs() yields the absolute value, and erfcinv() is the inverse complementary error function. We consider any value of $\sigma_P > 3.5$ to indicate significantly similar periods.

If the current TCE is (1) in a system that has a previous TCE dispositioned as an FP due to a stellar eclipse, (2) matches the previous TCE's period with $\sigma_P > 3.5$, and (3) is separated in phase from the previous TCE by at least 2.5 times the transit duration, then the current TCE is considered to be a secondary eclipse. In this case, it is designated as an FP and is classified into both the not transit-like and stellar eclipse FP categories — a unique combination that can be used to identify secondary eclipses while still ensuring they are not assigned *Kepler* Object of Interest numbers (see §6). Note that since the *Kepler* Pipeline generally identifies TCEs in order of their signal-to-noise, from high to low, sometimes a TCE identified as a secondary can have a deeper depth than the primary, depending on their relative durations and shapes. Also note that it is possible that the periods of two TCEs will meet the period matching criteria, but be different enough to have their relative phases shift significantly over the ≈4 year mission duration. Thus, the potential secondary TCE is actually required to be separated in phase by at least 2.5 times the previous TCE's transit duration over the entire mission time frame in order to be labeled as a secondary. Also, the *Kepler* Pipeline will occasionally detect the secondary eclipse of an eclipsing binary at a half, third, or some smaller integer fraction of the orbital period of the system, such that the epoch of the detected secondary coincides with that of the primary. Thus, when a non-1:1 period ratio is detected, we do not impose criteria (3), the phase separation requirement. Note, equations A1-A3 allow for integer period ratios.

A.3. *Not Transit-Like*

A very large fraction of false positive TCEs have light curves that do not resemble a detached transiting or eclipsing object. These include quasi-sinusoidal light curves from pulsating stars, star spots, and contact binaries, as well as more sporadic light curves due to instrumental artifacts. The first step in the catalog process is to determine whether each TCE is not transit-like. All transit-like obsTCEs are given *Kepler* Object of Interest (KOI) numbers, which are used to keep track of transit-like systems over multiple *Kepler* Pipeline runs. We employ a series of algorithmic tests to reliably identify these not transit-like FP TCEs, as shown by the flowchart in Figure 16.

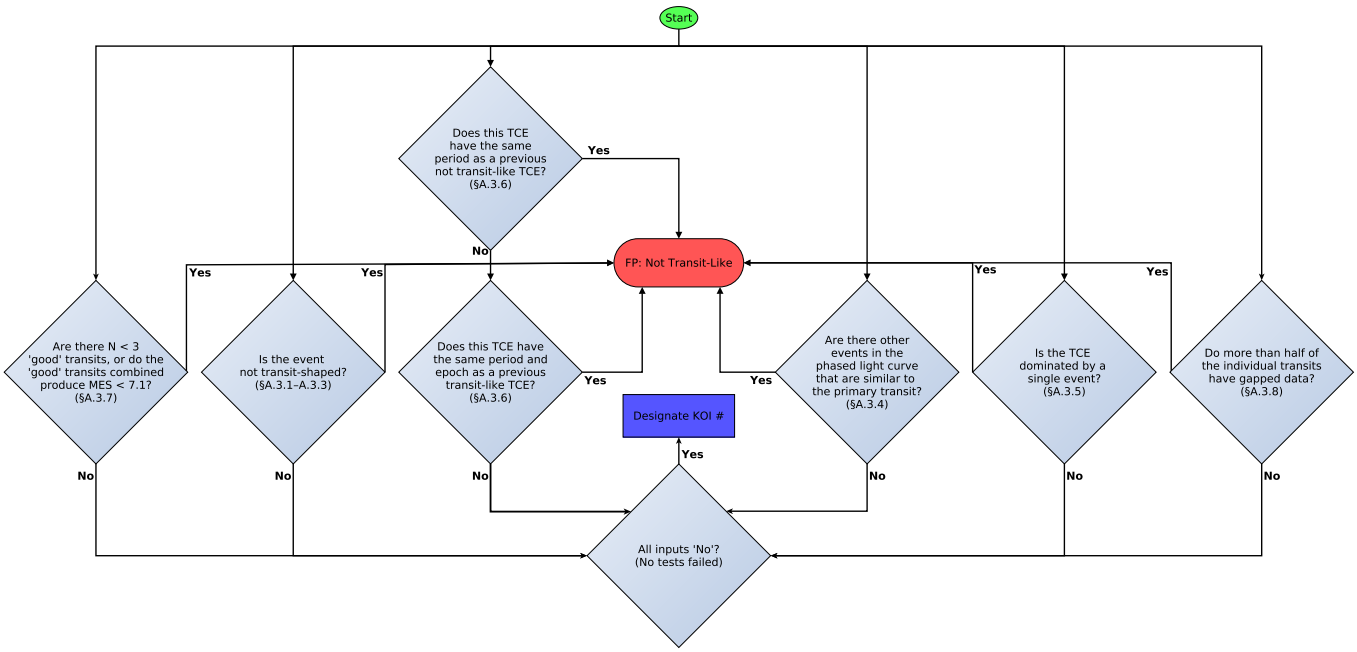

Figure 16. The not transit-like flowchart of the Robovetter. Diamonds represent "yes" or "no" decisions that are made with quantitative metrics. If a TCE fails any test (via a "yes" response to any decision) then it is dispositioned as a not transit-like FP. If a TCE passes all tests (via a "no" response to all decisions), then it is given a KOI number and passed to the stellar eclipse module (see §A.4 and Figure 21). The section numbers on each decision diamond correspond to the sections in this paper where these tests are discussed.

A.3.1. *The LPP Metric*

Many short-period FPs are due to variable stars that exhibit a quasi-sinusoidal phased light curve. We implement the LPP transit-like metric described by Thompson et al. (2015) to separate those TCEs that show a transit shape from those that do not. This technique bins the TCE's folded light curve and then applies a dimensionality reduction algorithm called Locality Preserving Projections (LPP, He & Niyogi 2004). It then measures the average Euclidean distance in these reduced dimensions to the nearest known transit-like TCEs to yield a single number that represents the similarity of a TCE's shape to that of known transits.

For the DR25 KOI catalog, we deviated slightly from the method described by Thompson et al. (2015)[22]. The DR24 LPP metric algorithm, when applied to DR25, produced LPP values that were systematically higher for short-period, low-MES TCEs. The transit duration of short period TCEs can be a significant fraction of the orbital period, so when folded and binned these transits have a noticeably different shape. And since we use injTCEs as our training set, which has very few short-period examples, there are very few known transits for the algorithm to match to, causing large measured distances for these transit event. The trend with MES is rooted in the fact that when the binned light curve has a lower signal-to-noise, it is less likely for two folded light curves to be similar to each other, creating more scatter in the reduced dimensions, and thus increasing the measured distance to known transits in those dimensions.

We reduced these dependencies by altering how we calculate the LPP metric for the DR25 KOI catalog. For our set of known transit-like TCEs, we now use the union of the set of recovered injTCEs and the set of PCs from the DR24 KOI catalog (Coughlin et al. 2016) that were re-found as obsTCEs in DR25. Including these PCs provides more examples at short period. We also changed how the folded light curve was binned. TCEs with lower MES are given wider bins for those cadences near the transit center, while keeping the total number of bins fixed (99 bins total including 41 for the in-transit portion). Finally, we divide these raw LPP values by the 75^{th} percentile of the raw LPP values for the 100 TCEs that are closest in period. In this way we reduce the period dependence in the LPP metric. Generally, the resulting LPP metric values lie near to a value of one, and values greater than ≈ 2 appear to be not-transit shaped. To create the DR25 catalog the Robovetter adopted a threshold of 2.2 for the DV detrending and 3.2 for the ALT detrending.

A.3.2. *Sine Wave Event Evaluation Test*

On occasion, a variable star's variability will have been mostly removed by both the DV and ALT de-

[22] The code is available here https://sourceforge.net/p/lpptransitlikemetric/

trendings and will thus appear transit-like. To identify these cases we developed the Sine Wave Event Evaluation Test (SWEET) to examine the PDC data and look for a strong sinusoidal signal at the TCE's period.

SWEET begins with the PDC data and normalizes each quarter by dividing the time series by the median flux value and subtracting 1.0. Outliers are robustly removed by utilizing a criterion based on the median absolute deviation (MAD) — specifically, outliers are identified as any point that lies more than $\sqrt{2}\cdot\mathrm{erfcinv}(1/N_{\mathrm{dat}})\sigma$ from the median, where N_{dat} is the number of data points, erfcinv is the inverse complementary error function, and 1σ=1.4826·MAD (see Hampel 1974; Ruppert 2010). Three different sine curves are fitted to the resulting data, with their periods fixed to half, exactly, and twice the TCE period, with their phase, amplitude, and offset allowed to vary. Of the three fits, the one with the highest signal-to-noise ratio, defined as the amplitude divided by its error, is chosen as the strongest fit. If a TCE has a SWEET signal-to-noise ratio greater than 50, an amplitude greater than the TCE transit depth in both the DV and ALT detrendings, and has a period less than 5.0 days, it fails as not transit-like.

A.3.3. *TCE Chases*

In §A.3.7.3 we describe a individual transit metric called Chases that assesses the detection strength of individual transit events relative to other signals nearby in time. TCE Chases takes the median value of these individual transit measurements. When the median value is less than 0.8 the TCE fails as not-transit-like. As with the individual Chases metric, TCE Chases is only calculated when the TCE has five or fewer transit events contributing to the signal. With more than five transit events, the individual transit events are not expected to be statistically significant, and the assumptions of the Chases metric no longer apply.

A.3.4. *The Model-Shift Uniqueness Test*

If a TCE under investigation is truly a PC, there should not be any other transit-like events in the folded light curve with a depth, duration, and period similar to the primary signal, in either the positive or negative flux directions, i.e., the transit event should be unique in the phased light curve. Many FPs are due to noisy, quasi-periodic signals (see §2) and thus are not unique in the phased light curve. In order to identify these cases, we developed a "model-shift uniqueness test" and used it extensively for identifying false positives in the Q1–Q12 (Rowe et al. 2015b), Q1–Q16 (Mullally et al. 2015), and DR24 (Coughlin et al. 2016) planet candidate catalogs.

See §3.2.2 of Rowe et al. (2015b) and page 23 of Coughlin (2017a) for figures and a detailed explanation of the "model-shift uniqueness test". Briefly, after removing outliers, the best-fit model of the primary transit is used as a template to measure the best-fit depth at all other phases. The deepest event aside from the primary (pri) transit event is labeled as the secondary (sec) event, the next-deepest event is labeled as the tertiary (ter) event, and the most positive (pos) flux event (i.e., shows a flux brightening) is labeled as the positive event. The significances of these events (σ_{Pri}, σ_{Sec}, σ_{Ter}, and σ_{Pos}) are computed assuming white noise as determined by the standard deviation of the light curve residuals. Also, the ratio of the red noise (at the timescale of the transit duration) to the white noise (F_{Red}) is computed by examining the standard deviation of the best-fit depths at phases outside of the primary and secondary events.

When examining all events among all TCEs, assuming Gaussian noise, the minimum threshold for an event to be considered statistically significant is given by

$$FA_1 = \sqrt{2}\cdot\mathrm{erfcinv}\left(\frac{T_{\mathrm{dur}}}{P\cdot N_{\mathrm{TCEs}}}\right) \tag{A4}$$

where T_{dur} is the transit duration, P is the period, and $\mathrm{N_{TCEs}}$ is the number of TCEs examined. (The quantity P/T_{dur} represents the number of independent statistical tests for a single target.) When comparing two events from the same TCE, the minimum difference in their significances in order to be considered distinctly different is given by

$$FA_2 = \sqrt{2}\cdot\mathrm{erfcinv}\left(\frac{T_{\mathrm{dur}}}{P}\right) \tag{A5}$$

We compute the following quantities to use as decision metrics:

$$MS_1 = FA_1 - \sigma_{\mathrm{Pri}}/F_{\mathrm{Red}} \tag{A6}$$

$$MS_2 = FA_2 - (\sigma_{\mathrm{Pri}} - \sigma_{\mathrm{Ter}}) \tag{A7}$$

$$MS_3 = FA_2 - (\sigma_{\mathrm{Pri}} - \sigma_{\mathrm{Pos}}) \tag{A8}$$

In the Robovetter, we disposition a TCE as a not transit-like FP if either $MS_1 > 1.0$, $MS_2 > 2.0$, or $MS_3 > 4.0$ in the DV detrending, or if either $MS_1 > -3.0$, $MS_2 > 1.0$, or $MS_3 > 1.0$ in the ALT detrending. These criteria ensure that the primary event is statistically significant when compared to the systematic noise level of the light curve, the tertiary event, and the positive event, respectively. We also fail TCEs as not transit-like if σ_{Pri} exactly equals zero in both the DV and ALT detrendings. A value of zero indicates that the fit failed for both detrendings, and suggests that something is fundamentally flawed with the TCE.

A.3.5. *Dominated by Single Event*

The depths of individual transits of planet candidates should be equal to each other, and thus assuming constant noise levels, the SNR of individual transits should be nearly equivalent as well. In contrast, most of the long-period FPs that result from three or more equidistant systematic events are dominated in SNR by one

event. The *Kepler* Pipeline measures detection significance via the Multiple Event Statistic (MES), which is calculated by combining the Single Event Statistic (SES) of all the individual events that comprise the TCE — both the MES and SES are measures of SNR. Assuming all individual events have equal SES values,

$$\mathrm{MES} = \sqrt{N_{\mathrm{Trans}}} \cdot \mathrm{SES} \tag{A9}$$

where N_{Trans} is the number of transit events that comprise the TCE. Thus, SES/MES = 0.577 for a TCE with three transits, and less for a greater number of transits. If the largest SES value of a TCE's transit events, $\mathrm{SES}_{\mathrm{Max}}$, divided by the MES is much larger than 0.577 (regardless of the number of transits), this indicates that one of the individual events dominates when calculating the SNR.

In the Robovetter, for TCEs with periods greater than 90 days, if $\mathrm{SES}_{\mathrm{Max}}/\mathrm{MES} > 0.8$ it is dispositioned as a not transit-like FP. The period cutoff of 90 days is applied because short-period TCEs can have a large number of individual transit events, which dramatically increases the chance of one event coinciding with a large systematic feature, thus producing a large $\mathrm{SES}_{\mathrm{Max}}/\mathrm{MES}$ value despite being a valid planetary signal.

A.3.6. *Previous TCE With Same Period*

Most quasi-sinusoidal FPs produce multiple TCEs at the same period, or at integer ratios of each other. If a TCE in a system has been declared as not transit-like due to another test, it is logical that all subsequent TCEs in that system at the same period, or ratios thereof, should also be dispositioned not transit-like. Thus, we match the period of a given TCE to all previous not transit-like FPs via equations A1-A3. If the current TCE has a period match with $\sigma_P > 3.25$ to a prior not transit-like FP, it is also dispositioned as a not transit-like FP.

Similarly, some TCEs are produced that correspond to the edge of a previously identified transit-like TCE in the system. This often results when the previous TCE corresponding to a transit or eclipse is not completely removed prior to searching the light curve for another TCE. Thus, we match the period of a given TCE to all previous transit-like TCEs via equations A1-A3. If the current TCE has a period match with $\sigma_P > 3.25$ to a prior transit-like FP, and the two epochs are separated in phase by less than 2.5 transit durations, the current TCE is dispositioned as a not transit-like FP. For clarity, we note that it is sometimes possible that the periods of two TCEs will meet the period matching criteria, but be different enough to have their epochs shift significantly in phase over the ∼4 year mission duration. Thus, if they are separated in phase by less than 2.5 transit durations at any point in the mission time frame, the current TCE is dispositioned as a not transit-like FP.

A.3.7. *Individual Transit Metrics*

A new approach implemented in DR25 is to examine individual transit events for each TCE and determine if they are transit-like. After rejecting these "bad" transit events, we check if either

- There are less than 3 "good" events left
- The re-computed MES using only 'good' events is < 7.1

If either of these conditions are met, then the TCE is failed as not transit-like. This is in line with the *Kepler* mission requirement of at least three valid transit events with a MES ≥ 7.1 in order to generate a TCE. In the following subsections we list the various tests we apply to each individual transit event.

A.3.7.1. *Rubble – Missing Data*—A number of TCEs from the *Kepler* Pipeline are based on transit events that are missing a significant amount of data either in-transit or just before and/or after. These tend to be false positives that are triggering on edges of gaps, or cases were a large amount of data has been removed and a TCE is being created from the residuals of previous TCEs in the system. We thus devised the Rubble metric to clean-up these fragments from the TCE list. The Rubble value for each individual transit is computed by dividing the number of *Kepler* cadences that are available in the DV time series by the number of cadences expected across two transit durations given *Kepler*'s regular 29.42 min cadence and the transit duration provided by the DV fit. If the Rubble value for the transit falls below threshold, then that transit is not counted as a valid transit. We adopted a threshold value of 0.5 to generate the DR25 KOI Catalog.

A.3.7.2. *Marshall – Transit Shape*—In the DR24 KOI Catalog, Coughlin et al. (2016) used the Marshall algorithm (Mullally et al. 2016) to identify and reject false alarm TCEs caused by short period transients in the data. Marshall fits the proposed transit with models of various transients and uses a Bayesian Information Criterion (BIC) to decide which model is the best explanation for the data. Simulations in Mullally et al. (2016) showed that Marshall was 95% complete for TCEs with periods > 150 days and correctly rejected 66% of simulated artifact events. The limit on Marshall's effectiveness at eliminating false alarms was that it used a parabola to describe the out-of-transit flux, which failed to capture much of the real observed stellar variability. To ensure high completeness, Marshall was tuned to prevent a variable continuum from causing true transits to be rejected, at the cost of a lower effectiveness.

For the DR25 KOI catalog, we use a Gaussian Process approach (GP, Rasmussen & Williams 2006) to provide an improved continuum model and increase our effectiveness, while maintaining our high completeness.

Briefly, our approach aims to model the covariance in the light curve to better fit the trends in our data. A similar approach was used by Foreman-Mackey et al. (2016) to model single transits due to very long period planets ($P > 1000$ days).

Our procedure is as follows. For each individual proposed transit event, we select a snippet of PDC data 30 times the reported transit duration centered on the event. Where the event happens near the start (or end) of a quarter, we take a snippet of similar length anchored at the start (or end) of the quarter. We use the *George* package (Ambikasaran et al. 2014) to fit the covariance of the out-of-transit flux with an exponential squared function, $\mathrm{Cov}(\delta t) = A \exp{(\delta t/\ell)^2}$, where A and ℓ are tunable parameters.

We next fit four models to the entire snippet.

$$\begin{aligned} G(t|A,\ell) + y_0 \\ G(t|A,\ell) + y_0 + S(t) \\ G(t|A,\ell) + y_0 + S(t)(1 - \exp \beta t) \\ G(t|A,\ell) + y_0 + S(t - \tau/2) - S(t + \tau/2) \end{aligned} \tag{A10}$$

where G is the Gaussian Process model with the tunable parameters held fixed to those found earlier, and y_0 is a constant offset. $S(t)$ is given by

$$S(t) = \frac{d}{1 + e^{-\gamma(t - t_0)}} \tag{A11}$$

where d and t_0 are tunable parameters and γ is a positive constant. This function, known as a sigmoid (or logistic) function, has asymptotes of 0 for $t << t_0$, and d for $t >> t_0$. The function transitions quickly, but smoothly, between the two states near $t = t_0$, where it takes on a value of d/2.

By using a sigmoid and avoiding the discontinuities present in the models used by the original Marshall algorithm (Mullally et al. 2016) we can use the L-BFGS-B algorithm (Byrd et al. 1995) available in the Scipy package [23] instead of the less robust Nelder-Mead.

The second function in equation A10 models a discrete jump in the data. We fit this model seeded with a negative-going dip at the predicted time of ingress, and also with a positive-going spike at the predicted egress, as we see both types of features in *Kepler* data. The third model fits a Sudden Pixel Sensitivity Drop (SPSD) event, probably caused by a cosmic ray hit on the detector. The last model approximates a box transit. By varying the parameter γ we could in principle model transit ingress and egress, but find that extra degree of freedom is not necessary to fit the low signal-to-noise events of most concern.

For each transit the Marshall method returns the BIC score, the preferred model, and the difference between the BIC scores of the preferred model and the sigmoid box fit. A transit is considered sufficiently bad when this difference (also known as the Marshall score) exceeds a particular threshold, as with the original Marshall algorithm. However, in a few cases the Gaussian process fails and yields extremely large, unbelievable BIC values. In these cases the transit is set to always pass. Also, for low MES transits, the expected SES of a transit is sufficiently low that Marshall will be unable to distinguish between the "no transit" model and a low signal-to-noise transit. Because of this the Robovetter declares a specific transit is not valid if all of the following criteria are met:

- The BIC score of the best-fitting non-transit model is at least 10 lower than the BIC of the transit-model
- The BIC score of the best-fitting non-transit model is less than 1.0E6
- Either $\mathrm{MES}/\sqrt{\mathrm{N_{RealTrans}}} > 4.0$ or the lowest BIC model is for the constant offset model,

Note, $\mathrm{N_{RealTrans}}$ is the total number of observed transit events for the TCE. The Marshall code used for the DR25 KOI catalog is available on sourceforge[24].

A.3.7.3. *Chases – SES artifacts*—The Chases metric was developed to chase-down non-transit like events on long period, low MES TCEs. Qualitatively, the metric mimics the human vetting preference to classify a TCE as a PC when individual transit events "stand-out" as a unique, transit-like signal from a visual inspection of the *Kepler* flux time-series data. In order to quantify this human vetting preference, we developed the the Chases algorithm. Chases uses the SES time series generated by the TPS module of the *Kepler* Pipeline (Jenkins 2017b). The SES time series measures the significance of a transit signal centered on every cadence. Details of calculating the SES time series is given in Jenkins et al. (2002) and illustrative examples are given in Tenenbaum et al. (2012). A transit produces a peak in the SES time series (as do systematic signals). TPS searches the SES time series for equally spaced peaks indicative of a series of transits. The series of individual peaks in the SES time series are combined to form the MES employed as the primary threshold for detecting a transit signal (Jenkins et al. 2002; Twicken et al. 2016; Jenkins 2017b).

The Chases metric quantifies how well the SES peaks contributing to a TCE approximate the expected shape and significance (relative to neighboring data) of a bona fide transit signal. Figure 17 shows the detrended flux time series (upper panel) and the corresponding SES time series (lower panel) for a clear single transit event

[23] www.scipy.org

[24] https://sourceforge.net/projects/marshall/

contributing to the TCE detection of K03900.01 on target KIC 11911580. The flux time series, with a very clear decrement during in-transit cadences (orange points), has the archetypal SES time series of a strong central peak with two low-amplitude, symmetric side troughs (caused by the way TPS uses wavelets to modify the model transits when calculating the SES, see Jenkins 2017b).

The Chases metric for an individual transit event is formulated by identifying the maximum SES value for cadences in transit, $SES_{\rm max}$ (in Figure 17, $SES_{\rm max} \approx 20$). Next, excluding cadences within $1.5\tau_{\rm dur}$ of mid transit (to avoid the symmetric side troughs), where $\tau_{\rm dur}$ is the detected transit duration, the SES time series is searched for Δ_t, the temporally closest feature to mid transit in the absolute value of the SES time series, $|SES|$. A feature is defined as when $|SES| > f\,SES_{\rm max}$, where f represents a tunable fraction of the peak in the SES time series. Finally, we define a maximum window $\Delta_{tmax} = P_{orb}/10$ with which to search for a comparable peak in $|SES|$, and form the final Chases metric for an individual transit event as $C_i = \min(\Delta_t, \Delta_{tmax})/\Delta_{tmax}$.

A value of $C_i = 1$ indicates that there is no comparable peak/trough in the SES time series within f of $SES_{\rm max}$ over the interval Δ_{tmax} of the transit signal. Thus, $Ch_i = 1$ is consistent with a unique, transit-like signal. A value of $Ch_i \approx 0$ indicates that a comparable strength feature is present in the SES time series temporally close to the transit event, and is consistent with the human vetting tendency to dismiss such signals as spurious. Figure 18 shows an example of a spurious TCE detection on the target KIC 11449918. The target is on a detector suffering from elevated levels of the "rolling-band" image artifacts as described in §A.3.7.4. The neighboring peak of comparable strength in the SES time series would result in $Ch_i \approx 0$ for this individual transit event. The Chases metric is also sensitive to the shape of the transit signal as illustrated in Figure 19. The SPSD shown in Figure 19 is a spurious instrumental signal with an asymmetric shape. Because Chases uses the absolute value of the SES, $Ch_i \approx 0$ for these types of events.

For each TCE with five or fewer transit events contributing to the signal, Ch_i is calculated for every transit event. With more than five transit events, the individual transit events are not expected to be statistically significant, and the assumptions of the Chases metric no longer apply. The individual transit event Ch_i values were used to recalculate the MES (see §A.3.7). Transit events with $Ch_i < 0.01$ were excluded from the Robovetter's MES calculation.

A.3.7.4. *Skye – Image Artifacts Clustered by Skygroup*—As discussed in 2.1, there are a number of TCEs caused by rolling-band image artifacts. These artifacts are caused by a spatial pattern in the CCD bias level that moves across the chip in response to changes in the temperature of the chip (for more detail see Van Cleve & Caldwell 2009). If a number of individual transit events from TCEs on different targets, but the same skygroup (region of the sky that falls on the same CCD each quarter), occur at the same time, they are very likely systematic in origin. The metric called Skye looks for an excess in the number of individual events occurring at the same time in the same skygroup. If an excess is identified we consider these events to be caused by artifacts.

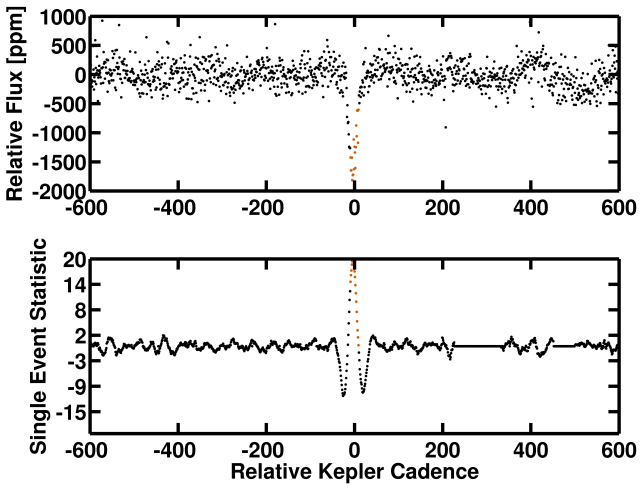

Figure 17. Upper panel: flux time series for a single transit event contributing to the TCE for KOI 3900.01 on target KIC 11911580 (black points). The cadences in transit (orange points) show a significant flux decrement relative to the baseline flux level. Lower panel: SES time series of the transit event show in the upper panel, representing the archetypal shape of a transit signal displaying a strong central peak with two low-amplitude, symmetric side troughs. There are no other events as strong as the transit nearby in time so this signal has an individual transit event Chases metric, $Ch_i = 1$.

More specifically, for each skygroup we bin the individual events into 1.0 d bins. We only use those obsTCE with periods greater than 45 d (∼half a *Kepler* quarter) for each skygroup. The reason for the period cut is that the long-period obsTCEs are likely to be affected by rolling-band systematics, but the short-period ones are not. Including shorter period TCEs would dramatically increase the number of individual transits and would reduce the significance of the anomalous peaks. See Figure 20 for an example of the anomalous peaks seen in some skygroups when the data is binned in this way.

To determine which events are anomalous, for each skygroup, we compute the average rate (R) of transits, by dividing the overall number of individual transit events in the skygroup by the number of 1.0 d bins. Assuming the majority of transits are randomly distributed

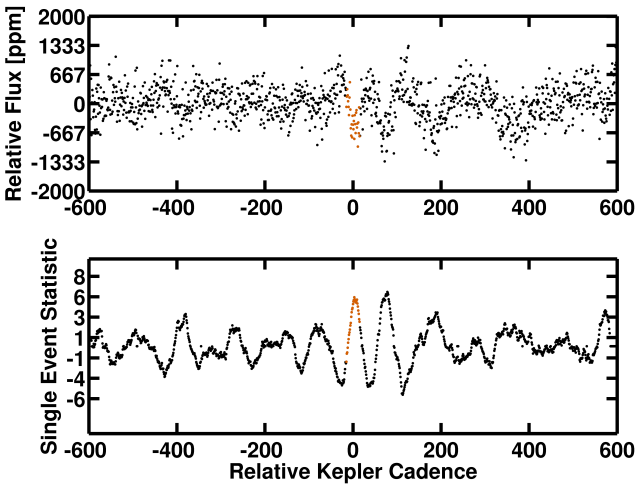

Figure 18. Upper panel: flux time series for a single transit event contributing to the TCE on target KIC 11449918 (black points). The cadences in transit (orange points) show a flux decrement, but there are numerous other flux decrements of similar depth and shape. The instrumental "rolling band" pattern noise contributes systematics to the flux time series of target KIC 11449918 causing numerous signal detections. Lower panel: SES time series of the transit event shown in the upper panel, representing the non-unique nature of the SES peak relative to surrounding data. The neighboring peak of comparable strength in the SES time series would result in $Ch_i = .016$ and the transit would be considered "bad" by Chases.

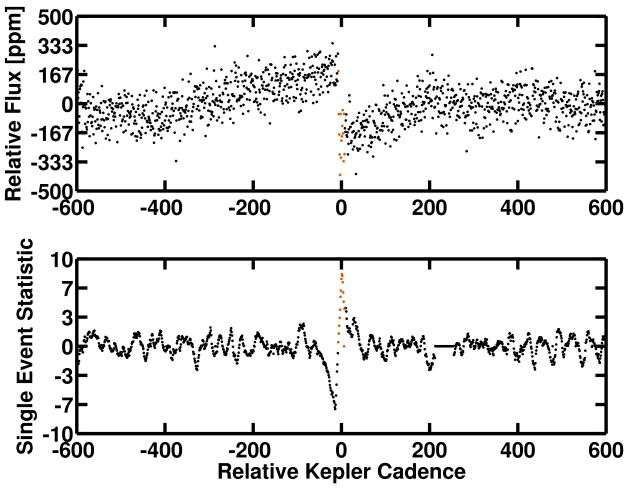

Figure 19. Upper panel: flux time series for a single transit event contributing to the TCE on target KIC 12357074 (black points). The cadences in transit (orange points) show a flux decrement, but the sudden drop in flux followed by the gradual return to the baseline is archetypal of the SPSD instrumental signature. Lower panel: SES time series for the transit event shown in the upper panel, illustrating the strongly asymmetric SES peak having a comparable amplitude negative SES trough preceding the SES peak. The neighboring trough of comparable absolute strength to the transit's peak would result in $Ch_i = .005$ and the transit would be considered "bad" by Chases.

in time, and utilizing Poisson counting statistics, any peaks greater than:

$$\text{threshold} = R + N \cdot \sqrt{R} \tag{A12}$$

are statistically significant and indicative of temporal clustering, given a chosen value for N. We choose a value of $N = 3.0$, and robustly determine the rate for each skygroup by first computing the threshold using all the bins, then iteratively rejecting all bins with a height greater than threshold and re-computing threshold until it converges and does not change with further iterations.

For each skygroup and its threshold, we identify the individual times-of-transit for TCEs belonging to the skygroup that fall in bins that are above the threshold. We assign Skye a value of 1.0 to these individual transits to indicate they are bad transits. The Skye value for all other transit times are set to zero.

A.3.7.5. *Zuma – Negative Significance*—A valid transit-like TCE should be comprised of individual events that correspond to flux decrements. If any event instead shows an increase of flux then that event is suspect. We thus designate any individual transit event with SES < 0 as "bad".

A.3.7.6. *Tracker – Ephemeris Slip*—After the TPS module of the *Kepler* Pipeline detects a TCE, it is sent to DV to be fit with a full transit model. DV allows the period and epoch to vary when fitting in order to provide as accurate a fit as possible. Sometimes the TPS ephemeris and DV ephemeris can end up significantly different. When this occurs it indicates that the underlying data is not transit-like and the TCE is likely due to quasi-sinusoidal systematics, which cause the ephemeris to wander when fitting.

Tracker measures (i.e., keeps track of) the time difference between the TPS and DV linear ephemerides in units of the TCE's duration for each transit. When Tracker is greater than 0.5 $T_{\rm dur}$ for any transit we designate the transit as bad.

A.3.8. *Fraction of Gapped Events*

Due to the method of data gapping employed in TPS, sometimes the *Kepler* Pipeline can create a TCE that has a majority of its individual events occur where there is no actual in-transit data. This tends to happen particularly in multi-TCE systems, because once the *Kepler* Pipeline detects a TCE in a given system, it removes the data corresponding to the in-transit cadences of that TCE, and re-searches the light curve.

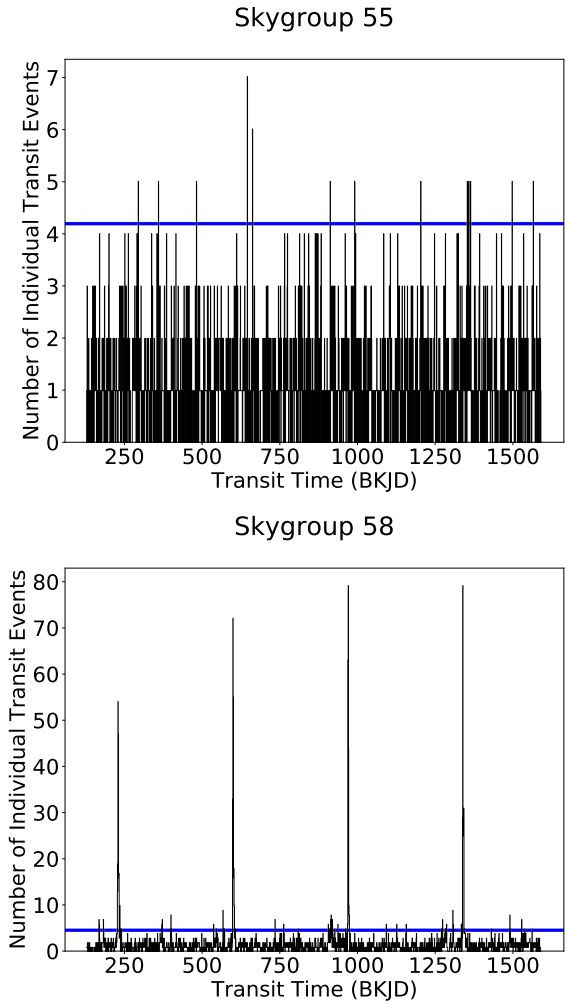

Figure 20. An example of how the Skye metric flags individual transit events. The plots show the number of individual transit events (from TCEs with periods greater than 45 days) that occur in one-day time bins throughout the mission duration. Two of the 84 skygroups were chosen to be shown as examples, with skygroup 55 plotted on top, and skygroup 58 plotted on bottom. Skygroup 58 (lower panel) has a strong clustering of transit events at times that correspond to the ∼372 day orbital period of the spacecraft, as the stars belonging to skygroup 58 fall on CCD channels with strong rolling-band signal. In contrast, skygroup 55 is nearly uniform. Individual transits that occur in a one-day time bin with a number of transit events above the threshold (shown by the blue horizontal line; see Equation A12) are flagged as bad transits due to the Skye metric.

We thus measure the number of individual transit events that actually contain data. Specifically, we compute the fraction of individual events with either SES $\neq 0$ or Rubble > 0.75, which indicate there is sufficient in-cadence data present. If the fraction of transits meeting these criteria is ≤ 0.5, we fail the TCE as not transit-like and give it the flag TRANS_GAPPED.

A.3.9. *No Data Available*

In a very small number of cases, neither the DV nor the ALT detrending produces a light curve and model fit for a TCE. This happens when the TCE is extremely not transit-like, usually due to a combination of severe systematics and a lack of substantial in-transit data. As a result, if no data from either detrending is available, the Robovetter fails a TCE as not transit-like.

A.4. *Stellar Eclipse*

If a TCE is deemed transit-like by passing all of the tests presented in §A.3 on both detrendings, it is given a KOI number (see flowchart in Figure 16). However, many of these KOIs are FPs due to eclipsing binaries and contamination from nearby variable stars. We employ a series of robotic tests to detect systems that are due to stellar companions, as shown by the flowchart in Figure 21.

A.4.1. *Secondary Eclipse*

One of the most common methods to detect a stellar system is the presence of a significant secondary in the light curve. With the exception of some hot Jupiter type planets (e.g., HAT-P-7, Borucki et al. 2009), the visibility of a secondary eclipse in *Kepler* data is a telltale sign of a stellar eclipsing binary.

A.4.1.1. *Subsequent TCE With Same Period*—Once the *Kepler* Pipeline detects a TCE in a given system, it removes the data corresponding to this event and researches the light curve. It is thus able to detect the secondary eclipse of an eclipsing binary as a subsequent TCE, which will have the same period, but different epoch, as the primary TCE. Thus, using equations A1-A3, the Robovetter dispositions a TCE as a stellar system FP if its period matches a subsequent TCE within the specified tolerance ($\sigma_P > 3.25$) and they are separated in phase by at least 2.5 times the transit duration. For clarity, we note again that it is sometimes possible that the periods of two TCEs will meet the period matching criteria, but be different enough to have their epochs shift significantly in phase over the ∼4 year mission duration. The phase separation requirement must be upheld over the entire mission duration in order to disposition the TCE as an FP due to a stellar eclipse.

Occasionally the *Kepler* Pipeline will detect the secondary eclipse of an eclipsing binary at half, third, or some smaller integer fraction of the orbital period of the system. In these cases, the epoch of the TCE corresponding to the secondary will overlap with that of the primary. These cases are accounted for by not requiring a phase separation of at least 2.5 transit durations when a period ratio other than unity is detected. (Note that equations A1-A3 allow for integer period ratios.) While

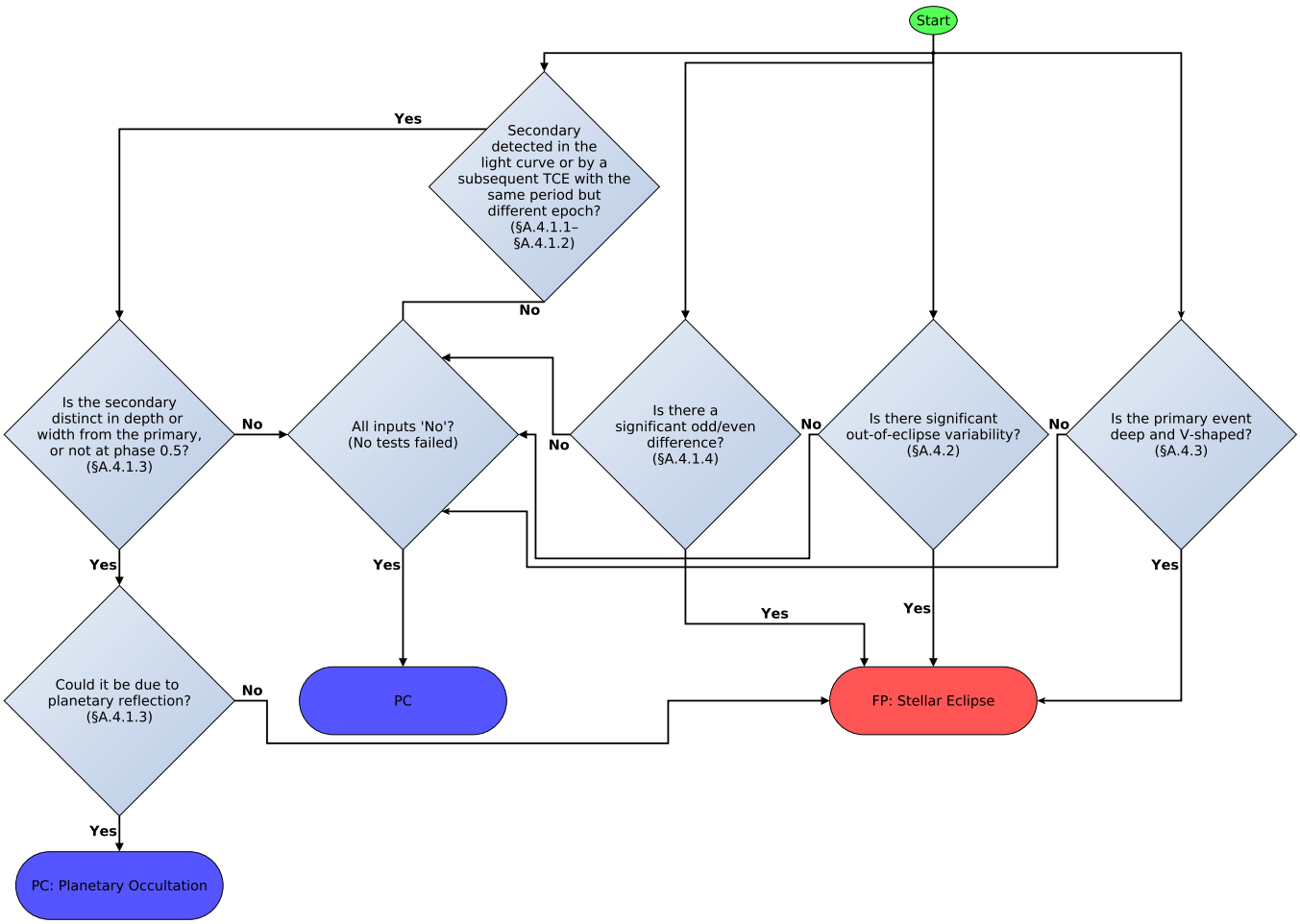

Figure 21. Flowchart describing the stellar eclipse tests of the Robovetter. Diamonds represent "yes" or "no" decisions that are made with quantitative metrics. The multiple arrows originating from "Start" represent decisions that are made in parallel.

this approach will likely classify any multi-planet system in an exact 2:1 orbital resonance as an FP due to a stellar eclipse, in practice this is non-existent. Exact 2:1 orbital resonances, where "exact" means the period ratio is close enough to 2.0 over the ∼4 year mission duration to avoid any drift in relative epoch, appear to be extremely rare (Fabrycky et al. 2014). Also, they might produce strong transit timing variations, which would likely preclude their detection. The *Kepler* Pipeline employs a strictly linear ephemeris when searching for TCEs, and thus while planets with mild transit timing variations (TTVs), e.g., deviations from a linear ephemeris less than the transit duration, are often detected, planets with strong TTVs, e.g., deviations from a linear ephemeris greater than the transit duration, are often not detected.

A.4.1.2. *Secondary Detected in Light Curve*—There are many cases when a secondary eclipse does not produce its own TCE, most often when its MES is below the *Kepler* Pipeline detection threshold of 7.1. The model-shift uniqueness test, discussed in §A.3.4, is well-suited to automatically detect secondary eclipses in the phased light curve, as it searches for the next two deepest events aside from the primary event. It is thus able to detect the best-candidate secondary eclipse in the light curve and assess its significance. We compute the following quantities to use as secondary detection metrics

$$MS_4 = \sigma_{\rm Sec}/F_{\rm Red} - FA_1 \tag{A13}$$

$$MS_5 = (\sigma_{\rm Sec} - \sigma_{\rm Ter}) - FA_2 \tag{A14}$$

$$MS_6 = (\sigma_{\rm Sec} - \sigma_{\rm Pos}) - FA_2 \tag{A15}$$

Recall that σ indicates a significance and was defined in §A.3.4. If MS_4 >1, MS_5 >0, and MS_6 >0, in either the DV or alternate detrendings, the Robovetter dispositions the TCE as a stellar system FP. These criteria ensure that the secondary event is statistically significant when compared to the systematic noise level of the light curve, the tertiary event, and the positive event, respectively.

A.4.1.3. *Candidates with Stellar Eclipses*—There are two exceptions when the above-mentioned conditions are met, but the Robovetter does not designate the TCE as an FP. First, if the primary and secondary widths and depths are statistically indistinguishable, and the secondary is located at phase 0.5, then it is possible that the TCE is a PC that has been detected at twice the true orbital period. Thus, the Robovetter labels a TCE with a stellar eclipse as a PC when $\sigma_{\rm Pri} - \sigma_{\rm Sec} < FA_2$ and the phase of the secondary is within 1/4 of the primary transit's duration of phase 0.5. Second, hot Jupiter PCs can have detectable secondary eclipses due to planetary occultations via reflected light and thermal emission (Coughlin & López-Morales 2012; Christiansen et al. 2010). Thus, a TCE with a detected stellar eclipse is labeled as a PC with the stellar eclipse flag (in order to facilitate the identification of hot Jupiter occultations) when the geometric albedo required to produce the observed secondary eclipse is less than 1.0, the planetary radius is less than 30 $R_\oplus$, the depth of the secondary is less than 10% of the primary, and the impact parameter is less than 0.95. The additional criteria beyond the albedo criterion are needed to ensure that this test is only applied to potentially valid planets and not grazing eclipsing binaries. We calculate the geometric albedo by using the stellar mass, radius, and effective temperature from the DR25 stellar catalog (Mathur et al. 2017), and the values of the period and radius ratio from the original DV fits.

A.4.1.4. *Odd/Even Depth Difference*—If the primary and secondary eclipses of eclipsing binaries are similar in depth, and the secondary is located near phase 0.5, the *Kepler* Pipeline may detect them as a single TCE at half the true orbital period of the eclipsing binary. In these cases, if the primary and secondary depths are dissimilar enough, it is possible to detect it as an FP by comparing the depths of the odd- and even-numbered transit events and their associated uncertainties, via the following statistic,

$$\sigma_{\rm OE} = \frac{abs\left(d_{\rm odd} - d_{\rm even}\right)}{\sqrt{\sigma_{odd}^2 + \sigma_{even}^2}} \tag{A16}$$

where $d_{\rm odd}$ is the measured depth using the odd-numbered transits, with associated uncertainty σ_{odd}, $d_{\rm even}$ is the measured depth using the even-numbered transits, with associated uncertainty σ_{even}, and abs() returns the absolute value.

We use two different methods to compute $d_{\rm odd}$, σ_{odd}, $d_{\rm even}$, σ_{even}, and thus $\sigma_{\rm OE}$, for both for the DV and ALT detrending. For the first method, the depths are computed by taking the median of all the points near the center of all transits, and the uncertainty is the standard deviation of those points, both using only the odd- or even-numbered transits. For the ALT detrending with a trapezoidal fit, we use all points that lie within ±30 minutes of the central time of transit, as well as any other points within the in-transit flat portion of the trapezoidal fit. For the DV detrending, we use all points within ±30 minutes of the central time of transit. (This threshold corresponds to the long-cadence integration time of the *Kepler* spacecraft. Including points farther away from the central time of transit degrades the accuracy and precision of the test.) If $\sigma_{\rm OE} > 1.1$ for either the DV or ALT detrending then the TCE is labeled as an FP due to a secondary eclipse and given the DEPTH_ODDEVEN_DV and/or DEPTH_ODDEVEN_ALT flag(s). The value of 1.1 was empirically derived using manual checks and transit injection. This method is very robust to outliers and systematics, but not extremely sensitive as it does not take into account the full transit shape to measure the depth.

The second method measures the depths and uncertainties by running the model-shift test separately on the portions of the light curve within half a phase of the odd- and even-numbered transits. Model-shift measures the depths and associated uncertainties using the entire transit model and taking into account the measured noise level of the entire light curve. This method is more sensitive to small odd/even differences, but also more sensitive to outliers and light curve systematics compared to the above method. If $\sigma_{\rm OE} > 11.2$ for the DV detrending, or > 19.8 for the ALT detrending, then the TCE is labeled as an FP due to a stellar eclipse and given the MOD_ODDEVEN_DV and/or MOD_ODDEVEN_ALT flag(s). The thresholds of 11.2 and 19.8 were empirically derived using manual checks and transit injection. This method is susceptible to outliers and systematics (and why the thresholds are set fairly high), but can also detect small, yet significant odd/even differences that the other method listed above cannot.

A.4.2. *Out of Eclipse Variability*

Short-period eclipsing binaries will often show out-of-eclipse variability due to tidal forces that deform the star from a perfect spheroid. The variability manifests as quasi-sinusoidal variations at either the period, or half the period, of the binary.

We use the information from SWEET (see §A.3.2) to detect these cases. If a transit-like TCE has a SWEET SNR greater than 50, an amplitude less than the TCE transit depth in either the DV and ALT detrendings, an amplitude greater than 5,000 ppm, and a period less than 10 days, we fail it as a stellar system.

A.4.3. *V-Shape Metric*

There are cases of eclipsing binaries that do not show a secondary eclipse, either due to the secondary star being too low luminosity for the eclipse to be detectable, or the binary has significant eccentricity and a longitude of periastron such that geometrically no eclipse occurs. Also, most detached eclipsing binaries will not exhibit detectable out-of-eclipse variability. In these cases, the only remaining way to infer that the signal is due to a

stellar system and not a planet is to utilize the shape and depth of the transit.

In previous catalogs (Rowe et al. 2015a; Mullally et al. 2015; Coughlin et al. 2016) TCEs were not failed based on their inferred radii alone. This was deliberate as the catalogs attempted to be as agnostic to stellar parameters as possible, such that dispositions would remain applicable if and when better stellar parameters were obtained, e.g., by GAIA (Cacciari 2009; Mignard 2005). This resulted in some PC KOIs with large depths that were known to very likely be eclipsing binaries, and in fact were later confirmed as such by follow-up observations (Santerne et al. 2016).

In this catalog, we attempt to strike a balance between identifying these binary systems, while still remaining agnostic to stellar parameters. We adapted a simple shape parameter, originally proposed in Batalha et al. (2013), and express it as the sum of the modeled radius ratio and the impact parameter. This metric reliably identifies eclipsing binaries both due to being too deep (large $R_p/R_\star$) and due to grazing eclipses (large impact parameter, b). Specifically we fail a transit-like TCE as a stellar system if $R_p/R_\star + b > 1.04$.

A.5. *Centroid Offset*

A.5.1. *Centroid Robovetter*

The Robovetter relies on a piece of code called the Centroid Robovetter[25] (Mullally 2017) to detect when a transit signal originates from a background or nearby star instead of from the target star. The Centroid Robovetter has not changed since its implementation for the DR24 KOI catalog; we summarize it below for completeness.

Given that *Kepler*'s pixels are 3.98″ square (Koch et al. 2010), and the typical photometric aperture has a radius of 4–7 pixels (Bryson et al. 2010), it is quite common for a given target star to be contaminated by light from another star. If that other star is variable, then that variability will be visible in the target aperture at a reduced amplitude. If the variability due to contamination results in a TCE, then it is a false positive, whether the contaminator is an eclipsing binary, planet, or other type of variable star (Bryson et al. 2013). For example, if a transit or an eclipse occurs on a bright star, a shallower event may be observed on a nearby, fainter star. Similarly, a star can be mistakenly identified as experiencing a shallow transit if a deep eclipse occurs on a fainter, nearby source.

The DV module of the *Kepler* Pipeline produces difference images for each quarter, which are made by subtracting the average flux in each pixel during each transit from the flux in each pixel just before, and after, each transit (Bryson et al. 2013). If the resulting difference image shows significant flux at a location (centroid) other than the target, then the TCE is likely an FP due to a centroid offset.

In our robotic procedure to detect FPs due to centroid offsets, we first check that the difference image for each quarter contains a discernible stellar image and is not dominated by background noise. This is done by searching for at least 3 pixels that are adjacent to each other and brighter than a given threshold, which is set by the noise properties of the image. We use an iterative sigma clipping approach to eliminate bright pixels when calculating the background noise, as the star often dominates the flux budget of a substantial number of pixels in the aperture.

For the difference images that are determined to contain a discernible stellar image, we first search for evidence of contamination from sources that are resolved from the target. Since resolved sources near the edge of the image may not be fully captured, attempts to fit models of the stellar profile often fail to converge. Instead, we check if the location of the brightest pixel in the difference image is more than 1.5 pixels from the location of the target star. If at least two-thirds of the quarterly difference images show evidence of an offset by this criterion, we disposition the TCE as an FP due to a centroid offset.

If no centroid offset is identified by the previous method, we then look for contamination from sources that are unresolved from the target. We fit a model of the pixel response function (PRF) to the difference images and search for statistically significant shifts in the centroid with respect to the PRF centroid of the out-of-transit images, or the catalog position of the source. Following Bryson et al. (2013), a TCE is marked as an FP due to a centroid offset if there are at least three difference images with a discernible stellar image, and a 3σ significant offset larger than 2″, or a 4σ offset larger than 1″ is measured.

The Centroid Robovetter gives the *Kepler* Robovetter several flags to indicate whether a centroid offset was detected and whether that detection can be trusted. The names of those flags have been changed for DR25 to be consistent with our minor flag naming scheme. A list of the minor flags are available in Appendix B.

A.5.2. *Ghost Diagnostic*

The last method we use to detect a centroid offset is the ghost diagnostic, which was added to the DR25 *Kepler* Pipeline (see § 11.3.7 of Jenkins 2017b). It determines whether a transit signal is likely contamination from a ghost image of a star located away from the target star in the focal plane. Ghost reflections occur when light from a bright star is reflected off the CCD and again from the field flattener plate and back onto the CCD. It appears as a diffuse, out-of-focus image of the pupil of the telescope. A similar type of false positive results from direct PRF (Pixel Response Function) con-

[25] https://sourceforge.net/projects/keplercentroidrobovetter/

tamination, when flux from the broad wings of a bright star near the target star on the CCD overlaps the target star's PRF. If a ghost reflection (or the PRF of a nearby star) containing a transit-like signature (e.g. an eclipsing binary signal) overlaps the PRF of the target star, then the contaminating transit signal will be equally strong in the periphery and the core of the target.

To detect this type of false alarm, the ghost diagnostic essentially measures the strength of the TCE signal in two separate light curves — one created using the average of the pixels inside the target's optimal aperture minus the average of the pixels in an annulus surrounding the target aperture (core aperture correlation statistic), and the other using the average of the pixels in the annulus surrounding the target aperture (halo aperture correlation statistic). If the ratio of the halo aperture to core aperture statistic is greater than 4.0, the TCE is marked as an FP with the major flag set to Centroid Offset. This ghost diagnostic is not available to vet the scrTCEs and thus the reliability measured with that set of TCEs will be too small by an insignificant amount.

A.6. *Ephemeris Matching*

Another method for detecting FPs due to contamination is to compare the ephemerides (periods and epochs) of TCEs to each other, as well as other known variable sources in the *Kepler* field. If two targets have the same ephemeris within a specified tolerance, then at least one of them is an FP due to contamination. Coughlin et al. (2014) used Q1–Q12 data to compare the ephemerides of KOIs to each other and eclipsing binaries known from both *Kepler*- and ground-based observations. They identified over 600 FPs via ephemeris matching, of which over 100 were not known as FPs via other methods. They also identified four main mechanisms of contamination. The results of Coughlin et al. (2014) were incorporated in Rowe et al. (2015b, see §3.3), and with some small modifications to Mullally et al. (2015, see §5.3) and Coughlin et al. (2016).

We modified the matching criteria used in previous catalogs to improve performance. We use the results of the transit injection run (§2.3) to measure the ability of the original DV fits by the *Kepler* Pipeline to recover period and epoch as a function of period. (Note that while the DV fits do produce an error on the measured period, it is not a robustly measured error, and thus not sufficient for our purposes.) In Figure 22 we show, in the top two panels, the difference in the injected and recovered period and epoch, as a function of the injected period. The bottom panels show the measured standard deviation of the difference as a function of period, in linear and logarithmic space respectively. The red line is the result of a best-fit power law.

When comparing two objects, A and B, where A is defined to have the shorter period, the new matching metrics we use, S_P and S_T for period and epoch respectively, are:

$$S_P = \frac{|P_r \cdot P_A - P_B|}{\sqrt{2} \cdot \sigma_P(P_A)} \tag{A17}$$

$$S_T = \frac{|T_A - T_B - T_r \cdot P_A|}{\sqrt{2} \cdot \sigma_T(P_A)} \tag{A18}$$

where P_A and P_B are the periods of objects A and B, T_A and T_B are similarly the epochs of objects A and B, $\sigma_P(P_A)$ and $\sigma_T(P_A)$ are the errors in period and epoch, given period P_A, derived from the best-fit power law to the standard deviation of the injected versus recovered periods and epochs, respectively. The period ratio, P_r, and epoch ratio, T_r, are defined by:

$$P_r = \text{rint}\left(\frac{P_B}{P_A}\right) \tag{A19}$$

$$T_r = \text{rint}\left(\frac{T_A - T_B}{P_A}\right) \tag{A20}$$

where rint() rounds a number to the nearest integer. Thus, a perfect match has $S_P = 0$ and $S_T = 0$, with worse matches having increasingly larger values of S_P and S_T.

We consider matches with $S_P < 5$ and $S_T < 5$, with period ratios of 50 or less ($P_r < 50$), to be statistically significant enough to constitute a match. We also require:

1. The two objects do not have the same KIC ID,
2. The two objects satisfy at least one of the following conditions:
   (a) A separation distance less than $d_{\max}$ arcseconds, where
   $$d_{\max}('') = 55 \cdot \sqrt{10^6 \cdot 10^{-0.4 \cdot m_{\rm kep}} + 1} \tag{A21}$$
   with the *Kepler* magnitude of the brighter source being used for $m_{\rm kep}$,
   (b) Located on opposite sides of the field-of-view center, but equidistant from the center to within a 100″ (25 pixel) tolerance.
   (c) Located on the same CCD module and within 5 pixels of the same column value in any of the 4 quarters.
   (d) Located on the same CCD module and within 5 pixels of the same row and column value in any of the 4 quarters.

Criterion 1 ensures that no star is ever matched to itself. Criterion 2a is a semi-empirically determined formula derived to account for direct PRF contamination and reflection off the field flattener lens, assuming the

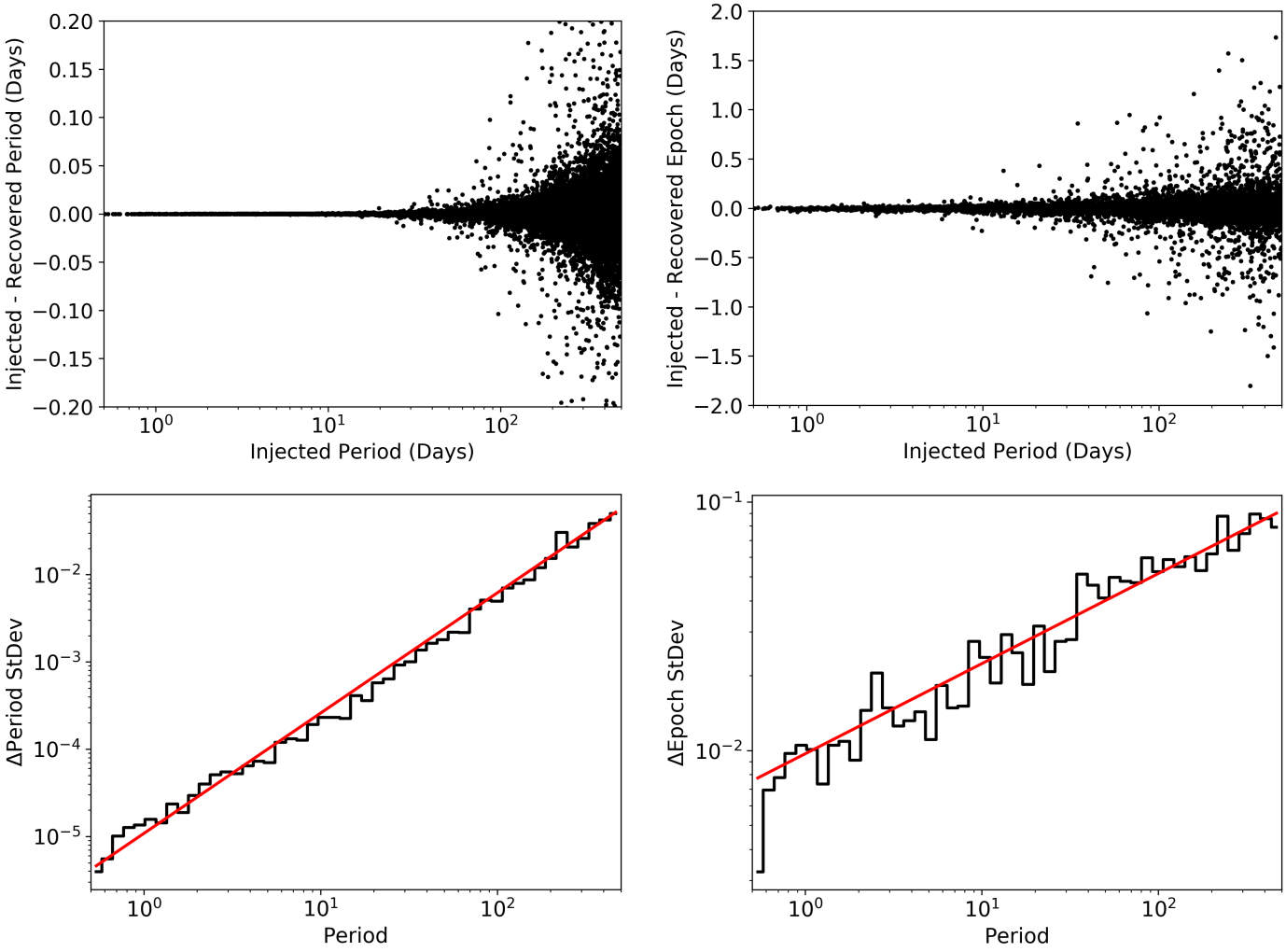

Figure 22. A plot of injected versus recovered periods and epochs of injected on-target planets. The top plots shows the difference between the injected and recovered periods (top) and epochs (right) as a function of period. The bottom plots show the measured standard deviation of the differences in period (left) and epoch (right) in logarithmic space. The red line shows the best-fit power-law in each case.

average wings of a *Kepler* PSF can be approximated by a Lorentzian distribution. The formula allows for any two stars to match within a generous 55″ range, but allows for bright stars to match to larger distances, e.g., a 10$^{\text{th}}$ mag star could match up to 550″ away, and a 5th mag star could match up to 5500″ away. Criterion 2b accounts for antipodal reflection off the Schmidt Corrector. Criterion 2c accounts for the column anomaly (see §3.5 of Coughlin et al. 2016), and criterion 2d accounts for video crosstalk.

In this Q1–Q17 DR25 catalog, we match the ephemerides of all Q1–Q17 DR25 TCEs (Twicken et al. 2016), including rogue TCEs, to the following sources:

- Themselves.
- The list of 8,826 KOIs from the NASA Exoplanet Archive cumulative KOI table after the closure of the Q1–Q17 DR24 table and publication of the last catalog (Coughlin et al. 2016).
- The *Kepler* Eclipsing Binary Working Group list of 2,605 "true" eclipsing binaries found with *Kepler* data as of 2016 October 13 (Prša et al. 2011; Slawson et al. 2011; Kirk et al. 2016).
- J.M. Kreiner's up-to-date database of ephemerides of ground-based eclipsing binaries as of 2016 October 13 (Kreiner 2004).
- Ground-based eclipsing binaries found via the TrES survey (Devor et al. 2008).
- The General Catalog of Variable Stars (GCVS Samus et al. 2009) list of all known ground-based variable stars, published 2016 October 05.

Via ephemeris matching, we identify 1,859 Q1–Q17 DR25 TCEs as FPs. Of these, 106 were identified as FPs only due to ephemeris matching. We list all 1,859 TCEs in Table 8, as this information is valuable for studying contamination in the *Kepler* field. In this table each TCE is identified by its KIC ID and planet number, separated by a dash. We also list in Table 8 each TCE's most likely parent, the period ratio between child and parent (P_{rat}), the distance between the child and parent in arcseconds, the offset in row and column between the child and parent in pixels (ΔRow and ΔCol), the magnitude of the parent (m_{Kep}), the difference in magnitude between the child and parent (ΔMag), the depth ratio of the child and parent (D_{rat}), the mechanism of contamination, and a flag to designate unique situations. In Figure 23 we plot the location of each FP TCE and its most likely parent, connected by a solid line. TCEs are represented by solid black points, KOIs are represented by solid green points, eclipsing binaries found by *Kepler* are represented by solid red points, eclipsing binaries discovered from the ground are represented by solid blue points, and TCEs due to a common systematic are represented by open black points. The *Kepler* magnitude of each star is shown via a scaled point size. Most parent-child pairs are so close together that the line connecting them is not easily visible on the scale of the plot.

Since *Kepler* does not observe every star in its field of view, it can often be the case that a match is found between two objects, but given their relative magnitude, distance, and depths it is clear that neither is the parent of the other, so these are classified as "bastards" (Coughlin et al. 2014). To identify the bastards due to direct PRF contamination, we performed a robust fit of the Kepler PRF model described by equations 9 and 10 of Coughlin et al. (2014) to the depth ratio, magnitude difference, and distance between each object identified as due to direct PRF contamination and its most likely parent. After iteratively rejecting outliers greater than 4.0 times the standard deviation, the fit converged with values of $\alpha = 6.93''$and $\gamma = 0.358''$. Outliers greater than 4.0 times the standard deviation of the final iteration, with these resulting fit parameters, were labeled as bastards. For the mechanism of column anomaly and reflection, if the depth ratio of the two objects is between 0.01 and 100, then it is labeled as a bastard, as these mechanisms should produce depth ratios of at least 1E-3 or 1E3. All bastards are identified with a flag of 1 in Table 8. Additionally, it can sometimes be the case that objects are matched via the column anomaly, but are on different outputs of the same module — these cases likely involve the column anomaly working in conjunction with cross-talk, and thus are complicated, and given a flag of 2 in Table 8. Finally, a flag of 3 indicates a combination of flags 1 and 2.

A.7. *Informational Only Tests*

There are a couple tests that the Robovetter performs that do not influence the disposition of a TCE. While failing one of these tests indicates a likely FP, it is not reliable enough to declare a TCE an FP. Instead, TCEs that fail these tests are given information only flags (see §B) as a way to notify users that a manual inspection of the TCE and the Robovetter results is likely warranted.

A.7.1. *Planet In Star*

In some cases, the DV fit returns a semi-major axis of the planetary orbit that is smaller than the radius of the host star. Such a fit is unphysical, as the planet would be orbiting inside the star; this is usually indicative of an FP. However, since many of the stellar parameters have large errors and their accuracy can vary, this situation does not guarantee the TCE is an FP. Thus, if a TCE is dispositioned as transit-like (the NT flag is not set), and if the semi-major axis from the DV fit is less than the stellar radius from the DR25 stellar properties catalog (Mathur et al. 2017), the TCE is flagged as PLANET_IN_STAR.

A.7.2. *Seasonal Depth Differences*

Due to the *Kepler* spacecraft's rotation every ≈90 days, each target and the surrounding stars will fall on a new CCD every quarter, and return to the same CCD once every four quarters. All of the quarters that correspond to the same CCD are labeled as being in a given season (e.g., Q2, Q6, Q10, and Q14 belong to Season 0, Q3, Q7, Q11, and Q15 belong to Season 1, etc., Thompson et al. 2016a). The shape and size of the optimal aperture for a given star is seasonally dependant and can change significantly season-to-season. As a result, a target will have differing amounts of third light in its optimal aperture from nearby stars. If the source of the signal that triggers a TCE is not from the target star, but rather from another source (as just discussed in §A.5 and §A.6), the level of contamination, and thus observed depth of the TCE, will have significant seasonal variation. Observation of seasonal depth differences is usually a good indication that the target is contaminated and a centroid offset is likely. However, depth differences can also arise when the signal is truly coming from the target, but significant third light exists in the aperture and the seasonal variations are not sufficiently corrected.

In order to automatically detect seasonal depth differences, if a TCE has been dispositioned as transit-like (the NT flag is not set), we measure the depth and associated error of the primary event in each season utilizing the first method described in the second paragraph of §A.4.1.4, i.e., we compute the median and standard deviation of all the points within ±15 minutes of the center of transit. We then obtain an average depth over all seasons, D_a, by computing the mean of the depths of all four seasons.

Table 8. The 1,859 Q1–Q17 DR25 TCEs Identified as FPs due to Ephemeris Matches

TCE	Parent	P_{rat}	Distance ($''$)	ΔRow (Pixels)	ΔCol (Pixels)	m_{Kep}	ΔMag	D_{rat}	Mechanism	Flag
001433962-01	3924.01	1:1	13.5	3	-2	14.91	0.56	4.7434E+02	Direct-PRF	0
001724961-01	001724968-01	1:1	4.7	1	-1	13.39	-2.96	2.1190E+00	Direct-PRF	0
002166206-01	3735.01	1:1	8.3	-1	-2	17.64	-4.34	5.6706E+02	Direct-PRF	0
002309585-01	5982.01	1:1	11.7	-2	1	13.93	1.45	2.0011E+02	Direct-PRF	0
002437112-01	3598.01	1:1	19.7	-5	1	17.63	-1.48	1.0525E+03	Direct-PRF	0
002437112-02	002437149-02	2:1	19.7	-5	1	17.63	-1.48	6.9253E+02	Direct-PRF	0
002437488-01	6268.01	1:1	10.6	0	3	16.98	-2.02	2.5330E+02	Direct-PRF	0
002437804-01	002437783-01	1:1	14.4	4	-1	17.30	-3.14	1.4225E+02	Direct-PRF	0
...	...	...	...	...	...	...	...	...	...	...

NOTE—A suffix of "pri" in the parent name indicates the object is an eclipsing binary known from the ground, and the child TCE matches to its primary. Similarly a suffix of "sec" indicates the child TCE matches the secondary of a ground-based EB. Parent names are listed, in priority order when available, by (1) their Bayer designation (e.g., RR-Lyr-pri), (2) their EBWG (Eclipsing Binary Working Group; Kirk et al. 2016) designation (e.g., 002449084-pri), (3) their KOI number (e.g., 3924.01), and (4) their TCE number (e.g., 001724968-01). A flag of 1 indicates that the TCE is a bastard, which are cases where two or more TCEs match each other via the Direct-PRF contamination mechanism, but neither can physically be the parent of the other via their magnitudes, depths, and distances, and thus the true parent has not been identified. A flag of 2 indicates cases of column anomalies that occur on different outputs of the same module. These cases likely involve cross-talk to carry the signal from one output to another. TCEs due to the common systematic do not have information listed for a parent source, as they are not caused by a single parent. Note that Table 8 is published in its entirety in the electronic edition of the Astrophysical Journal. A portion is shown here for guidance regarding its form and content.

The significance of the seasonal depth differences, $S_{\rm Diff}$, is then computed via,

$$S_{\rm Diff} = \frac{\sum_{n=0}^{3} |D_n - D_a|}{\sqrt{\sum_{n=0}^{3} \sigma_n^2 + N \cdot \sigma_a^2}} \tag{A22}$$

where n denotes a particular season (0, 1, 2, or 3), N is the total number of seasons with a measured depth and uncertainty, D_n is the measured depth in a given season, σ_n is the measured error on the depth in a given season, D_a is the measured averaged depth, and σ_a is the measured error of the average depth, given by,

$$\sigma_a = \frac{\sqrt{\sum_{n=0}^{3} \sigma_n^2}}{N} \tag{A23}$$

For either the DV or ALT detrending, if $S_{\rm Diff} > 3.6$ then the TCE is flagged as having significant seasonal depth differences via the flag SEASONAL_DEPTH_(ALT|DV).

A.7.3. *Period Aliasing*

In some cases, the *Kepler* Pipeline detects a signal (and produces a TCE) that is at an integer multiple of the signal's true period. In most cases, this is due to the presence of seasonal depth differences, as the Pipeline ends up only locking onto events in the quarters with the strongest (deepest) signal. While this usually indicates an FP due to a centroid offset, as discussed in A.7.2, it is not a definitive measure. Also, the Pipeline will detect real planets with significant TTVs at longer (near integer multiple) periods.

In order to detect a period alias, we utilize the Model-shift results — if the TCE's period is an integer multiple of the signal's true period, then several, equally spaced events should be visible in the phased light curve. If the TCE has been dispositioned as transit-like (the NT flag is not set), the Robovetter first checks if Model-shift detected significant secondary and tertiary events, by ensuring that $\sigma_{\rm Sec}/F_{\rm Red} > FA_1$ and $\sigma_{\rm Ter}/F_{\rm Red} > FA_1$. If so, the phases of the secondary and tertiary events, $\phi_{\rm Sec}$ and $\phi_{\rm Ter}$, are then expressed as the absolute value of the their distance in phase from the primary event, i.e., constrained to be between 0.0 and 0.5. (For example, if secondary and tertiary events were initially detected at phases of 0.1 and 0.7, then $\phi_{\rm Sec} = 0.1$ and $\phi_{\rm Ter} = 0.3$.) If period aliasing is present, then $\phi_{\rm Sec}$ and $\phi_{\rm Ter}$ should be $\approx n/N$, where N is the integer multiple of the true signal that the Pipeline detected it at, and n is an integer between 1 and $N-1$ that is different for the secondary and tertiary events. (E.g., in the case of $\phi_{\rm Sec} = 0.1$ and $\phi_{\rm Ter} = 0.3$, this implies $N = 10$, $n = 1$ for $\phi_{\rm Sec}$, and $n = 3$ for $\phi_{\rm Ter}$).

We derive metrics to measure how close $\phi_{\rm Sec}$ and $\phi_{\rm Ter}$ each are to an exact integer period alias, called $S_{\rm Sec}$ and $S_{\rm Ter}$. Specifically,

$$\begin{aligned} S_{\rm Sec} &= \sqrt{2} \cdot {\rm erfcinv}\left(\left|\frac{1}{\phi_{\rm Sec}} - {\rm rint}\left(\frac{1}{\phi_{Sec}}\right)\right|\right) \\ S_{\rm Ter} &= \sqrt{2} \cdot {\rm erfcinv}\left(\left|\frac{1}{\phi_{\rm Ter}} - {\rm rint}\left(\frac{1}{\phi_{Ter}}\right)\right|\right) \end{aligned} \tag{A24}$$

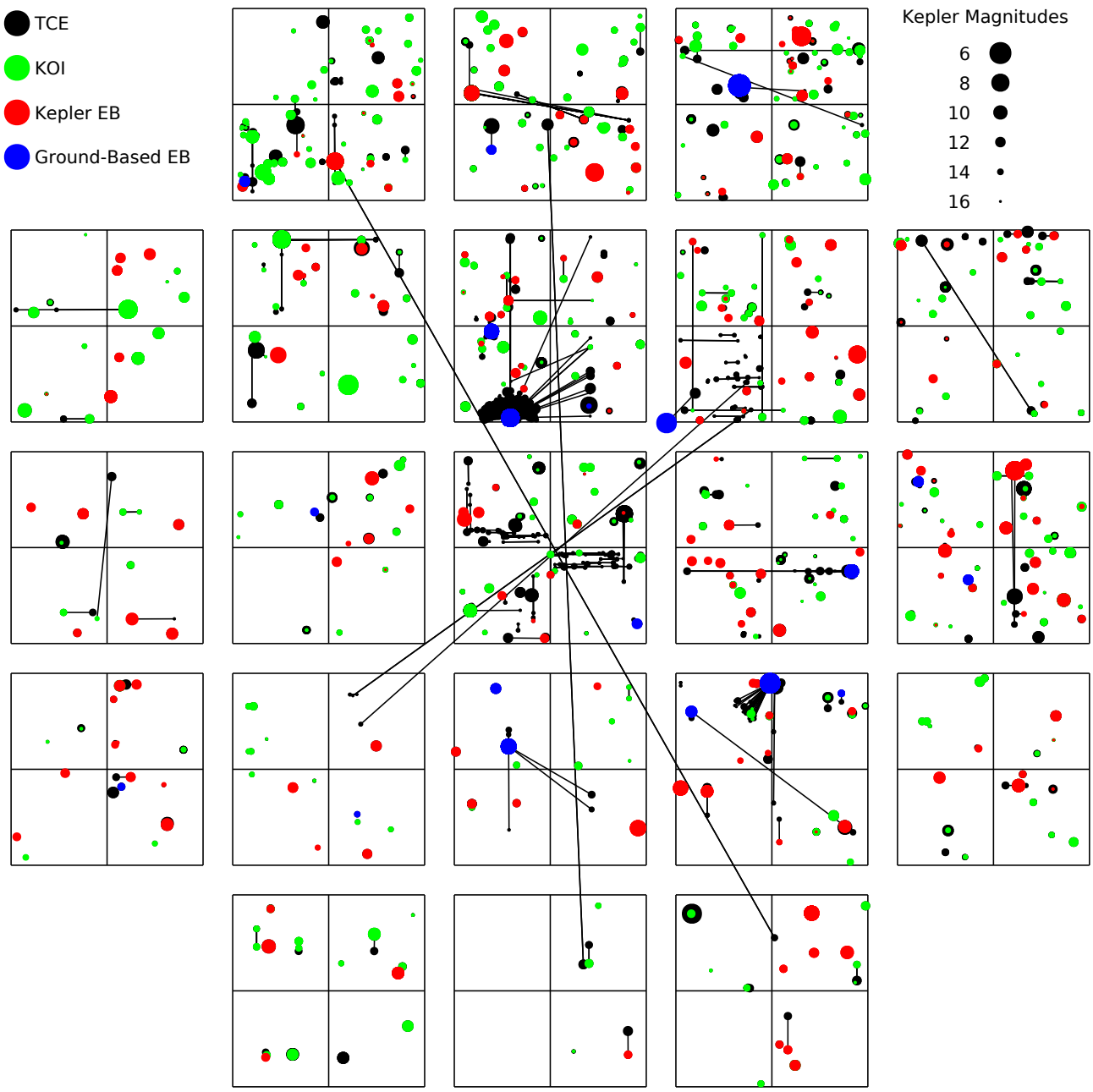

Figure 23. Distribution of ephemeris matches on the focal plane. Symbol size scales with magnitude, while color represents the catalog in which the contaminating source was found. Blue indicates that the true transit is from a variable star only known as a result of ground-based observations. Red circles are stars listed in the *Kepler* EBWG catalog (Kirk et al. 2016, http://keplerebs.villanova.edu/), green are KOIs, and black are TCEs. Black lines connect false positive matches with the most likely contaminating parent. In most cases parent and child are so close that the connecting line is invisible.

where erfcinv() is the inverse complementary error function, and rint() rounds a number to the nearest integer. The higher the values of $S_{\rm Sec}$ and $S_{\rm Ter}$, the more closely the measured phases of the significant secondary and tertiary events correspond to an integer period ratio. These computations are performed independently for the DV and ALT detrendings. If $S_{\rm Sec} > 2.0$ and $S_{\rm Ter} > 2.0$, for either detrending, the Robovetter considers a period alias detected, and the TCE is flagged as PERIOD_ALIAS_(ALT|DV).

B. MINOR FALSE POSITIVE FLAG DEFINITIONS

The Robovetter produces a flag each time it gives a disposition of FP, and sometimes when it gives a disposition of PC. Here we give a definition for each flag. Table reft:minorstats shows the number and percentage of obsTCEs (not including rogue and banned) that were flagged with each minor flag. These flags are available for the KOIs through the comment column in the KOI table at the Exoplanet Archive. See the Robovetter output files[26] for the flags for all the obsTCEs, injTCEs, invTCEs, scrTCEs. A summary of the Robovetter metrics is given in Table 3.

ALL_TRANS_CHASES: This flag is set when the per TCE Chases metric is above threshold. This indicates that the shapes of the individual transits are generally not reliable and the TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.3.

CENT_CROWDED: This flag is set as a warning that more than one potential stellar image was found in the difference image, and thus a reliable centroid measurement cannot be obtained. See §A.5.1.

CENT_FEW_DIFFS: Fewer than 3 difference images of sufficiently high SNR are available, and thus very few tests in the pipeline's centroid module are applicable to the TCE. If this flag is set in conjunction with the CENT_RESOLVED_OFFSET flag, it serves as a warning that the source of the transit may be on a star clearly resolved from the target. See §A.5.1.

CENT_FEW_MEAS: The PRF centroid fit used by the pipeline's centroid module does not always converge, even in high SNR difference images. This flag is set as a warning if centroid offsets are recorded for fewer than 3 high SNR difference images. See §A.5.1.

CENT_INVERT_DIFF: One or more difference images were inverted, meaning the difference image claims the star got brighter during transit. This is usually due to variability of the target star and suggests the difference image should not be trusted. When this flag is set, it is a warning that the TCE requires further scrutiny, but the TCE is not marked as an FP due to a centroid offset. See §A.5.1.

[26] The Robovetter output files have the format kplr_dr25_XXX_robovetter_output.txt (XXX represents the data set name) and can be found in the Robovetter github repository, https://github.com/nasa/kepler-robovetter

CENT_KIC_POS: This measured offset distance is relative to the star's recorded position in the Kepler Input Catalog (KIC), not the out of transit centroid. Both are useful, since the KIC position is less accurate in sparse fields, but more accurate in crowded fields. If this is the only flag set, there is no reason to believe a statistically significant centroid shift is present. See §A.5.1.

CENT_NOFITS: The transit was not fit by a model in DV and thus no difference images were created for use by the pipeline's centroid module, so this flag is set as a warning that the TCE cannot be evaluated. This flag is typically set for very deep transits due to eclipsing binaries. See §A.5.1.

CENT_RESOLVED_OFFSET: The TCE has a significant centroid offset because the transit occurs on a star that is spatially resolved from the target. The TCE is marked as an FP with the centroid offset flag set unless one of the other Centroid Robovetter flags is also set, casting doubt on the measurement. See §A.5.1.

CENT_SATURATED: The star is saturated, so the Robovetter's centroiding assumptions break down. This flag is set as a warning, indicating that the TCE cannot be reliably evaluated. See §A.5.1.

CENT_UNCERTAIN: The significance of the centroid offset cannot be measured to high enough precision, so this flag is set as a warning that the TCE cannot be confidently dispositioned as an FP. This is typically due to having only a very small number (i.e., 3 or 4) of offset measurements, all with low SNR. See §A.5.1.

CENT_UNRESOLVED_OFFSET: There is a statistically significant shift in the centroid during transit. This indicates the is not on the target star. Thus, the TCE is dispositioned as an FP with the centroid offset major flag set, unless another Centroid Robovetter flag is also set, casting doubt on the measurement. See §A.5.1.

DEEP_V_SHAPED: The V-shape metric is above threshold. This metric uses the fitted DV radius ratio and impact parameter to determine whether the event is likely to be caused by a stellar eclipse. When the flag is set, the TCE is dispositioned as an FP with the stellar eclipse major flag set. See §A.4.3.

DEPTH_ODDEVEN_(ALT|DV): The TCE failed the odd-even depth test using the ALT or DV detrending. This determines whether the difference in the depths of the odd and even transits is greater than the standard deviation of the measured depths. The transit-like TCE is marked as an FP with a stellar eclipse major flag set. See §A.4.1.4.

EPHEM_MATCH: The TCE has been identified as an FP due to an ephemeris match with a source that could plausibly induce the observed variability on the target. See §A.6 and Table 8 for the contaminating source.

HALO_GHOST: The ghost diagnostic value is too high. This diagnostic measures the transit strength for

the out- and in-aperture pixels and determines if the transit is localized on the target star, or if it is due to contamination from a distant source. The TCE is an FP and the centroid offset major flag is set. See §A.5.2.

HAS_SEC_TCE: Another TCE on the same target with a higher planet number has the same period as the current transit-like TCE, but a significantly different epoch. This indicates that the current TCE is an eclipsing binary with the other TCE representing the secondary eclipse. If the PLANET_OCCULT_DV and PLANET_OCCULT_ALT flags are not set, the TCE is dispositioned as an FP with a stellar eclipse major flag set. See §A.4.1.1.

INCONSISTENT_TRANS: The ratio of the maximum SES value to the MES value is above threshold and the TCE has a period greater than 90 days. This flag indicates that the TCE has only a few transits and the MES is dominated by a single large event. Thus, the TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.5.

INDI_TRANS_(CHASES|MARSHALL|SKYE|ZUMA|TRACKER|RUBBLE): One or more of the individual transit metrics (Chases, Marshall, Skye, Zuma, Tracker, or Rubble) removed a transit causing the TCE's recalculated MES to drop below threshold, or the number of transits to drop below 3. The TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.7.

IS_SEC_TCE: The TCE has the same period, but a different epoch, as a previous transit-like TCE on the same target. This indicates that the current TCE corresponds to the secondary eclipse of an eclipsing binary (or a planet if the PLANET_OCCULT_DV or PLANET_OCCULT_ALT flags are set). Thus, the current TCE is dispositioned as an FP with both the not transit-like and stellar eclipse major flags set. See §A.2.

LPP_(ALT|DV): The Locality Preserving Projections (LPP) valueThompson et al. (2015), as computed using the ALT or DV detrending, is above threshold. This indicates that the TCE is not transit-shaped, and thus is dispositioned as an FP with the not transit-like major flag set. See §A.3.1.

MOD_NONUNIQ_(ALT|DV): The Model-shift 1 test, performed with the ALT or DV detrending, is below threshold. This test calculates the significance of the primary event, taking into account red noise, and compares it to the false alarm threshold. This flag indicates the primary event is not significant compared to the amount of systematic noise in the light curve, and thus the TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.4.

MOD_ODDEVEN_(ALT|DV): The odd/even statistic from the Model-shift test is calculated with the ALT or DV detrending. This statistic compares the best-fit transit model to the odd and even transits separately and determines that the difference in the resulting significance values is above threshold. When set, the transit-like TCE is dispositioned as an FP with the stellar eclipse major flag set. See §A.4.1.4.

MOD_POS_(ALT|DV): The Model-shift 3 test, performed with the ALT or DV detrending, is below threshold. This test compares the significance of the primary and positive-going events in the phased light curve to help determine whether the primary event is unique. This flag indicates that the TCE is likely noise and thus is dispositioned as an FP with the not transit-like major flag set. See §A.3.4.

MOD_SEC_(ALT|DV): The Model-shift 4, 5, and 6 values, calculated using the ALT or DV detrending, are above threshold. This test calculates the significance of the secondary event divided by F_{red}, the ratio of red noise to white noise in the light curve. The same calculation is done for the difference between the secondary and tertiary event significance values, and the difference between the secondary and positive event significance values. They indicate that there is a unique and significant secondary event in the light curve (i.e., a secondary eclipse). Thus, assuming the PLANET_OCCUL_(ALT|DV) flag is not set, the TCE is dispositioned as an FP with the stellar eclipse major flag set. See §A.4.1.2.

MOD_TER_(ALT|DV): The Model-shift 2 test, performed with the ALT or DV detrending, is below threshold. This test calculates the difference between the primary and tertiary event significance values. This flag indicates that the primary event is not unique in the phased light curve, and thus the TCE is likely noise and dispositioned as an FP with the not transit-like major flag set. See §A.3.4.

NO_FITS: Both the trapezoidal and the original DV transit fits failed to converge. This indicates the signal is not sufficiently transit-shaped in either detrending to be fit by a transit model. The TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.9.

PERIOD_ALIAS_(ALT|DV): Using the phases of the primary, secondary, and tertiary events from the Model-shift test run on the ALT or DV detrended data, a possible period alias is seen at a ratio of N:1, where N is an integer of 3 or greater. This indicates the TCE has likely been detected at a period that is N times longer than the true orbital period. This flag is currently informational only and not used to declare any TCE an FP. See §A.7.3.

PLANET_IN_STAR: The original DV planet fits indicate that the fitted semi-major axis of the planet is smaller than the stellar radius. As it is possible that the stellar data is not accurate, this flag is currently informational only and not used to declare any TCE an FP. See §A.7.1.

PLANET_OCCULT_(ALT|DV): A significant secondary eclipse was detected in the ALT or DV detrending, but it was determined to possibly be due to planetary reflection and/or thermal emission. While

the stellar eclipse major flag remains set, the TCE is dispositioned as a PC. See §A.4.1.3.

PLANET_PERIOD_IS_HALF_(ALT|DV): A significant secondary eclipse was detected in the ALT or DV detrending, but it was determined to be the same depth as the primary within the uncertainties. Thus, the TCE is possibly a PC that was detected at twice the true orbital period. When this flag is set, it acts as an override to other flags such that the stellar eclipse major flag is not set, and thus the TCE is dispositioned as a PC if no other major flags are set. See §A.4.1.3.

RESIDUAL_TCE: The TCE has the same period and epoch as a previous transit-like TCE. This indicates the current TCE is simply a residual artifact of the previous TCE that was not completely removed from the light curve. Thus, the current TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.6.

SAME_NTL_PERIOD: The current TCE has the same period as a previous TCE that was dispositioned as an FP with the not transit-like major flag set. This indicates that the current TCE is due to the same not transit-like signal. Thus, the current TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.6.

SEASONAL_DEPTH_(ALT|DV): There appears to be a significant difference in the computed TCE depth from different seasons using the ALT or DV detrending. This indicates significant light contamination, usually due to a bright star at the edge of the aperture, which may or may not be the origin of the transit-like event. As it is impossible to determine whether or not the TCE is on-target from this flag alone, it is currently informational only and not used to declare any TCE an FP. See §A.7.2.

SWEET_EB: The sine wave event evaluation test (SWEET) is above threshold, the detected signal has an amplitude less than the TCE's depth, and the TCE period is less than 5 days. This flag indicates that there is a significant sinusoidal variability in the PDC data at the same period as the TCE due to out-of-eclipse EB variability. The transit-like TCE is dispositioned as an FP with the stellar eclipse major flag set. See §A.4.2.

SWEET_NTL: The sine wave event evaluation test (SWEET) is above threshold, the detected signal has an amplitude greater than the TCE's depth, and the TCE period is less than 5 days. This flag indicates that there is a significant sinusoidal variability in the PDC data at the same period as the TCE, and the detected event is due to stellar variability and not a transit. The TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.2.

TRANS_GAPPED: The fraction of gapped transit events is above threshold. This flag indicates that a large number of observable transits had insufficient in-cadence data. The TCE is dispositioned as an FP with the not transit-like major flag set. See §A.3.8.

Table 9. obsTCEs Minor Flag Statistics

Minor Flag	Num. Flagged	% Flagged
ALL_TRANS_CHASES	8176	25.145
CENT_CROWDED	42	0.129
CENT_FEW_DIFFS	8957	27.547
CENT_FEW_MEAS	589	1.811
CENT_KIC_POS	1635	5.028
CENT_NOFITS	1952	6.003
CENT_RESOLVED_OFFSET	1956	6.016
CENT_SATURATED	3820	11.748
CENT_UNCERTAIN	89	0.274
CENT_UNRESOLVED_OFFSET	743	2.285
DEEP_V_SHAPED	895	2.753
DEPTH_ODDEVEN_ALT	220	0.677
DEPTH_ODDEVEN_DV	177	0.544
EPHEM_MATCH	1841	5.662
HALO_GHOST	3150	9.688
HAS_SEC_TCE	1141	3.509
INCONSISTENT_TRANS	7219	22.202
INDIV_TRANS_	14541	44.721
_CHASES	5468	16.817
_MARSHALL	7614	23.417
_SKYE	4790	14.732
_ZUMA	2103	6.468
_TRACKER	1880	5.782
_RUBBLE	7137	21.950
IS_SEC_TCE	1136	3.494
LPP_ALT	9948	30.595
LPP_DV	19271	59.268
MOD_NONUNIQ_ALT	11376	34.987
MOD_NONUNIQ_DV	11380	34.999
MOD_ODDEVEN_ALT	487	1.498
MOD_ODDEVEN_DV	401	1.233
MOD_POS_ALT	5578	17.155
MOD_POS_DV	4672	14.369
MOD_SEC_ALT	1407	4.327
MOD_SEC_DV	1161	3.571
MOD_TER_ALT	5340	16.423
MOD_TER_DV	4970	15.285
NO_FITS	113	0.348
PERIOD_ALIAS_ALT	5	0.015
PERIOD_ALIAS_DV	2	0.006
PLANET_IN_STAR	87	0.268
PLANET_OCCULT_ALT	18	0.055
PLANET_OCCULT_DV	39	0.120
PLANET_PERIOD_IS_HALF_ALT	18	0.055
PLANET_PERIOD_IS_HALF_DV	4	0.012
RESIDUAL_TCE	107	0.329
SAME_NTL_PERIOD	2061	6.339
SEASONAL_DEPTH_ALT	89	0.274
SEASONAL_DEPTH_DV	83	0.255
SWEET_EB	209	0.643
SWEET_NTL	1377	4.235
TRANS_GAPPED	5428	16.694

Note—For these statistics the obsTCE set does not include the rogue or banned TCEs. Most obsTCEs fail more than one test, so the percentages are not expected to add up to 100%.