



Scaling *K2*. II. Assembly of a Fully Automated *C5* Planet Candidate Catalog Using EDI-Vetter

Jon K. Zink^{1,2} , Kevin K. Hardegree-Ullman² , Jessie L. Christiansen² , Courtney D. Dressing³ , Ian J. M. Crossfield^{4,5} , Erik A. Petigura¹ , Joshua E. Schlieder⁶ , and David R. Ciardi² 

¹ Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA; jjzink@astro.ucla.edu

² Caltech/IPAC-NASA Exoplanet Science Institute, Pasadena, CA 91125, USA

³ Department of Astronomy, University of California, Berkeley, CA 94720, USA

⁴ Department of Physics, and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁵ Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA

⁶ Exoplanets and Stellar Astrophysics Laboratory, Code 667, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

Received 2019 October 8; revised 2020 January 27; accepted 2020 February 6; published 2020 March 13

Abstract

We present a uniform transiting exoplanet candidate list for Campaign 5 of the *K2* mission. This catalog contains 75 planets with seven multi-planet systems (five double, one triple, and one quadruple planet system). Within the range of our search, we find eight previously undetected candidates, with the remaining 67 candidates overlapping 51% of the study of Kruse et al. that manually vets candidates from Campaign 5. In order to vet our potential transit signals, we introduce the Exoplanet Detection Identification Vetter (EDI-Vetter), which is a fully automated program able to determine whether a transit signal should be labeled as a false positive or a planet candidate. This automation allows us to create a statistically uniform catalog, ideal for measurements of planet occurrence rate. When tested, the vetting software is able to ensure that our sample is 94.2% reliable against systematic false positives. Additionally, we inject artificial transits at the light-curve level of the raw *K2* data and find that the maximum completeness of our pipeline is 70% before vetting and 60% after vetting. For convenience of future studies of occurrence rate, we include measurements of stellar noise (CDPP) and the three-transit window function for each target. This study is part of a larger survey of the *K2* data set and the methodology that will be applied to the entirety of that set.

Unified Astronomy Thesaurus concepts: [Exoplanet catalogs \(488\)](#); [Planetary science \(1255\)](#); [Astronomy data analysis \(1858\)](#)

Supporting material: data behind figure, machine-readable tables

1. Introduction

The *Kepler* spacecraft was launched in 2009 and has found over 4000 transiting exoplanet candidates. The original mission lasted four years and collected photometric data on more than 150,000 stars in a 116 deg² patch of the sky (Koch et al. 2010; Borucki 2016). As intended, the fixed position and continuous (29.42 minute cadence) photometry provided the ideal conditions for transit detection and studies of occurrence rate (see, e.g., Youdin 2011; Howard et al. 2012; Dong & Zhu 2013; Dressing & Charbonneau 2013, 2015; Petigura et al. 2013; Burke et al. 2015; Mulders et al. 2015, 2018). With the final announcement of Data Release 25 (DR25; Thompson et al. 2018) and data from *Gaia* Data Release 2 (DR2; Gaia Collaboration et al. 2018), additional studies have been conducted using improved stellar parameters (Bryson et al. 2019; Hardegree-Ullman et al. 2019; Hsu et al. 2019; Zink et al. 2019a; Zink & Hansen 2019), providing a baseline measurement for the occurrence of Galactic exoplanets.

Upon the failure of two reaction wheels, the spacecraft was no longer able to remain focused on the same field for extended periods of time, thus concluding the original mission. However, the spacecraft was able to take photometric data despite the telescope's pointing issues. With the telescope now looking at different fields of the Galaxy for shorter periods of the time (<80 days), the *K2* mission was born (Howell et al. 2014; Van Cleve et al. 2016). The *K2* mission continued for 18 complete campaigns (fields) along the ecliptic, before the spacecraft

eventually ran out of fuel. Each campaign delivered a unique glimpse at a different part of the ecliptic plane, providing the opportunity to consider how these differences in Galactic latitude, metallicity, and stellar age (Rizzuto et al. 2018) may play a role in the occurrence of exoplanets. Making measurements of occurrence rate with *K2* will also allow us to combine data from *Kepler* to increase our sample size for global measurements of occurrence rate.

Several groups have constructed catalogs of different *K2* fields, yielding over 800 new exoplanet candidates (Adams et al. 2016; Barros et al. 2016; Crossfield et al. 2016; Pope et al. 2016; Vanderburg et al. 2016; Dressing et al. 2017; Livingston et al. 2018; Mayo et al. 2018; Petigura et al. 2018; Yu et al. 2018; Kruse et al. 2019; Zink et al. 2019b). Almost all of these catalogs have been focused on Campaigns 1–10, leaving a vast trove of unexamined data in Campaigns 11–18. Furthermore, vetting of the *K2* transit signals has been almost entirely done by eye. Many of the pointing issues the spacecraft experienced created artificial dips that mimic transit signals. *K2* data are considerably less well-behaved than the *Kepler* data as a result of this pointing jitter, leading to many systematic false positives (FPs) with high signal-to-noise ratio (S/N). Previous vetting software used for the *Kepler* catalogs, which were not tailored to these systematics, could have difficulty isolating these FPs, but they are easily detected by human examination. This lack of automation and repeatability, thus far, makes calculations of *K2* planet occurrence difficult to perform.

One notable attempt to automate the search of *K2* data was performed by Kostov et al. (2019), where previously detected candidates were vetted in a partly automated fashion (automated metrics prioritized candidates, which could then be verified with human examination). However, this method still requires some human interaction, making measurements of completeness and reliability difficult to achieve. Dattilo et al. (2019) made further progress in automating *K2* planet detections by using a convolutional neural network to classify signals. Despite high classification accuracy, the software still requires human supervision.

It is essential for any calculation of planet occurrence that we know, and can account for, the biases to which the empirical sample is subjected. The automated pipeline for *Kepler* was able to directly measure various metrics that can account for this lack of completeness (Mullally et al. 2015; Coughlin et al. 2016; Thompson et al. 2018). To test how good the pipeline is at recovering transit signals, previous studies have injected artificial transit signals into the light curves and measured the number of signals recovered by the automated pipeline (Christiansen et al. 2013, 2015; Christiansen 2017). This gave a direct measure of pipeline detection efficiency. Burke & Catanzarite (2017a) provided a measure of the target window function (the probability that a signal will provide three transits within the window of data available) and the 1σ depth function (which is a measure of the stellar noise given the transit duration). Both of these tools help us to calculate the probability of detecting a given transiting exoplanet. Without these metrics, strong assumptions about a catalog must be made in order to produce any measure of occurrence, and any conclusions should be taken with caution. Finally, to determine how pure the DR25 candidate sample was, reliability tests were performed by both inverting and scrambling the light curves and testing the pipeline for FPs (Thompson et al. 2018). Using a fully automated pipeline is the only way to achieve an accurate measure of systematic reliability.

Using the *Kepler* data set as a baseline detectable planet population, Dotson et al. (2019) estimates that *K2* should yield 1317 ± 261 detectable planets. However, the *Kepler* field may be intrinsically unique, distorting such estimates. With all the current *K2* candidates detected through non-automated vetting processes, it remains difficult to estimate what fraction of the potential candidates have been found. In this paper we provide a fully automated *K2* pipeline and candidate sample for Campaign 5. We select Campaign 5 for testing because this field has the largest known sample of *K2* planet candidates (246, as of 2019 October 7), allowing us to tune the vetting metrics to maximize our planet yield. This paper presents a methodology that is in the process of being applied to all the *K2* campaigns. In Section 2 we discuss how our pipeline takes the raw flux data and removes the noise from the spacecraft before searching for and finding transit-like signals. In Section 3 we introduce our fully automated vetting software, which has been optimized for *K2* systematics. In Section 4 we present our measure of stellar noise for each target. In Section 5 we inject artificial transit-like signals into the light curves in order to measure the completeness of our pipeline. In Section 6 we present the results of our pipeline reliability test. In Section 7 the results of our window function measurements are presented. In Section 8 we present our list of fully vetted candidates. In Section 9 we provide concluding remarks on our pipeline and candidate list.

2. Data Processing

In the following section we will discuss how our pipeline takes the raw photometric data and converts them into usable transit signals. In brief, the target pixel files are downloaded from MAST (Campaign 5, Data Releases 31) and processed with the EVEREST Python package (Luger et al. 2016, 2018). The light curves then undergo two Gaussian process (GP) regressions and a harmonic signal removal. Finally, the fully detrended data are examined for transit signals using the TERRA software (Petigura et al. 2013).

2.1. EVEREST

The *K2* data set provides unique systematics that make the detection of exoplanets more difficult than that of the original *Kepler* mission. The spacecraft thrusters and roll cause periodic fluctuations in the photometric data. In a preliminary study, we compared different *K2* detrending software: K2SFF (Vanderburg & Johnson 2014), K2PHOT (Petigura et al. 2013), and EVEREST⁷ (Luger et al. 2016, 2018). We found that all of these algorithms had different performance issues and strengths. The DAVE vetting software (Kostov et al. 2019) considers several different detrended light curves simultaneously for each target, but such a task is computationally expensive when considering the desire to directly execute each detrending algorithm. We found that the EVEREST software provided good minimization of signal rms and the most user-friendly software for removal of spacecraft fluctuation. In overview, EVEREST uses the calibrated pixel-level data to produce time-series photometric light curves and corrects for systematic noise in those light curves by examining pixel-level correlations. All the *K2* Campaigns have been processed with EVEREST (however, the calibrated pixel-level data files are currently being reprocessed as part of the *K2* Global Uniform Reprocessing Effort,⁸ warranting an updated processing with EVEREST). Light curves with EVEREST processing are publicly accessible.⁹ However, to ensure a uniform treatment, we undertook the task of reprocessing the Campaign 5 light curves with EVEREST, using the K2SFF aperture #15, as suggested by Luger et al. (2018). This aperture uses a model of the *Kepler* pixel response function to determine the size and shape of the target aperture (Bryson et al. 2010; Vanderburg & Johnson 2014). This gave us the ability to later inject signals into the raw flux data (at the light-curve level) before EVEREST processing (see Section 5). By directly running EVEREST on the calibrated pixel-level data (downloaded from MAST),¹⁰ we could ensure that the transit injection recovery is processed using the exact same conditions as those of the candidate sample.

On average, EVEREST masks about 8% of the light curve due to spacecraft anomalies or outliers. Occasionally, EVEREST will remove a large fraction of the data points within a light curve because the excessive noise is seen as numerous outliers. We remove targets where more than 50% of the light curve is masked. This severe masking only occurs in 37 of the Campaign 5 light curves, in most cases where the target is near the edge of the CCD. These 37 targets represent a very small fraction of the 25,040 light curves considered for this campaign.

⁷ <https://rodluger.github.io/everest/>

⁸ <https://keplerscience.arc.nasa.gov/k2-uniform-global-reprocessing-underway.html>

⁹ <https://archive.stsci.edu/hlsp/everest>

¹⁰ <https://archive.stsci.edu/k2/>

2.2. GP and Harmonic Detrending

The EVEREST software removes the systematics caused by the thrusters and roll of the spacecraft, but an additional global detrending is needed to flatten the light curves. We achieve this through GP regression and harmonic removal.

The first step in detrending is to establish the white noise level (σ). It is important that the detrending knows what is signal versus noise, and an accurate σ value will help to achieve this goal. In well-behaved flat light curves this is an easy task, and σ is the standard deviation of the normalized flux. However, if the time series contains correlated noise, the standard method will artificially increase the measured uncertainty. To complicate this even further, deep transit signals or a large number of outliers can again inflate the rms. We use the median absolute deviation (MAD), which is a measure of the distance of the median absolute value from the median value in the time series. This value can be scaled to estimate σ for the time series:

$$\sigma_{\text{MAD}} = \frac{\text{MAD}}{\Phi^{-1}(3/4)} \approx 1.4826 \text{ MAD} \quad (1)$$

where $\Phi^{-1}(3/4)$ is the normal inverse cumulative distribution function evaluated at a probability of 3/4. This measure of σ is far more robust to outliers and transit signals. However, light curves that require significant red-noise removal will still inflate the inferred σ value. To limit this issue we break the light curve into bins of width 1 day, because the daily flux trend should be minimal. We measure the σ_{MAD} value for each of these bins and take the median σ_{MAD} value from these bins as our measured value for σ . It is important to note that stellar variability on a scale of <1 day can still inflate our measured σ value, but this effect is insignificant when looking for long-term trends, as is done here.

One challenge of GP detrending is that GPs can fit real transit signals. A thorough discussion of this issue can be found in Hippke et al. (2019), with several alternative detrending methods which we hope to consider in future iterations of our pipeline. To avoid this issue, we attempt to mask outliers and potential transits. We again bin the time series by intervals of single days. Within each bin we look for points that exceed $3\sigma_{\text{MAD}}$ from the median of each bin. These points are only masked for the detrending process and unmasked for the rest of our pipeline. By manually adjusting the threshold, we found that a 3σ clipping was sufficient to mask most significant transits. Additionally, any transit existing below this limit was likely to be seen as noise by the GP and therefore not removed. An additional restriction we place on our GP is that we do not allow it to choose detrending periods <5 days. Periods shorter than this are more prone to overfitting transit signals because the transit duration is often a significant fraction of a day. Putting this limitation in place decreases our chances of modifying the transit depth. We discuss this issue more thoroughly in Section 8.

To implement our GP regression the pipeline uses PyMC3 (Salvatier et al. 2015) with a wrapper built by the Exoplanet software (Foreman-Mackey et al. 2019). Our detrending uses the ‘‘Rotation,’’ kernel which is a combination of two simple harmonic oscillators, meant to mimic stellar rotation/noise. Our pipeline uses this GP detrending in two steps: First, it looks for long-term trends (periods >10 days and general flux drifting), then after removing the long-term trends, it looks for

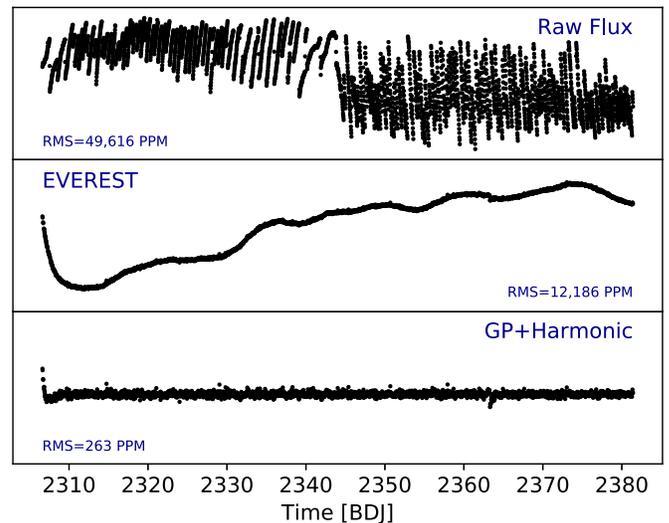


Figure 1. Top: the raw target pixel file data for EPIC 211422469. Middle: the data after processing with the EVEREST detrending software. Bottom: the light curve after an additional processing from our GP and harmonic fitter routine. The corresponding rms values are shown, in units of parts per million (ppm), to illustrate how each step will condense the noise of the light curve.

short-term trends ($5 \text{ days} < \text{period} < 10 \text{ days}$) and subtracts the best fit GP. However, stellar oscillations can also exist on very short timescales (often <0.5 days). Any attempt to remove these oscillations with a GP will almost certainly remove transits as well, thus a harmonic fitter is used to address this issue.

Unlike a GP, a harmonic fitter does not have the flexibility to fit and remove long-period transits. It can only affect the signals with strong similarities to a sine function, and the long periods of flat baseline between transits make this unlikely. This can become problematic when one has a system with multiple transiting planets, decreasing the gaps between transits and mimicking a sine curve (Christiansen et al. 2013; Zink et al. 2019a). Additionally, short-period giant planets can also be affected by this harmonic removal process, by either decreasing the transit depth or removing the signal completely. By choosing an upper limit of 0.5 days for our harmonic fitter, we minimize such occurrences. We begin by unmasking the GP σ -clipped data and searching for the strongest periodic signal below 0.5 days using a Lomb–Scargle periodogram. We then use a χ^2 fitting algorithm to measure the amplitude (A) of the signal. It is important that we do not introduce additional noise into the data set, so we require $A > 10\sigma_A$ before removing the harmonic signal. We found in practice that this limit rarely overfit the data, but we address such possibilities in our vetting metrics (Section 3.9). In Figure 1 we show how the raw flux data will change at each step of our light-curve processing.

Looking at our GP and harmonic fitter, there is an apparent gap in our considered period range for detrending ($0.5 \text{ days} < \text{period} < 5 \text{ days}$). Any attempt made to fit this period range either introduced additional artificial signals into our data or artificially reduced transit depths. Fortunately, most stars have rotational periods >5 days (McQuillan et al. 2013) and most variable stars oscillate at periods <0.5 days, indicating that few non-planetary periodic signals occupy this parameter space. Additionally, our vetting algorithm rejects any signals found with such harmonic features, minimizing the possibility of an FP detection (Section 3.9).

2.3. TERRA

Once the light curve has been fully processed, we begin our search for threshold crossing events (TCEs). These are signals with at least three transit-like events that reach or exceed a specific statistical significance. We discuss our process of selecting this threshold in the next Section (2.4). To search for these signals, we use the TERRA search algorithm (Petigura et al. 2013). This software begins by masking outliers and then uses a finely spaced grid-based search algorithm to find the event with the largest S/N in phase-folded period space of the light curve. This 3D grid search provides a measure of period (P), the transit duration (t_{dur}), the transit ephemeris (t_0), and S/N. For transiting planets detected by the *Kepler* pipeline, S/N is measured using the multiple event statistic (MES; Jenkins 2002), which indicates the strength of the whitened signal assuming a linear ephemeris. We will use MES in reference to the strength of the signal henceforth.¹¹

To recover systems with more than one transiting exoplanet, we consider the five signals with the largest MES in the light curve. We do not expect many systems to contain more than five distinct detectable transits within a period of 38 days (Kepler-80 and TRAPPIST-1 are currently the only known exceptions). However, Kruse et al. (2019) found evidence for one such system (EPIC 210965800) with six planets in Campaign 4 of *K2*. We acknowledge that we may lose some of these higher multiplicity systems by limiting our search to five planets, but such a restriction is necessary to minimize our computational cost. After the largest MES signal has been detected, we then mask $2.5\times$ the transit duration ($1.25t_{\text{dur}}$ on either side of the predicted transit midpoint), allowing the next largest signal to be detected. This process continues until either five TCEs have been found, or the largest MES value is less than the detection threshold. Zink et al. (2019a) showed that this type of masking will make higher multiplicity systems more difficult to detect, because nearby transits have the potential of being partially or fully hidden by this type of masking. However, removing the signal with a transit model is impractical. A majority of the signals are FPs that do not match well with the transit model. Thus, subtracting the best fit model will not actually remove the signal, but rather morph the data so that the grid search continues to detect the same anomaly repeatedly. In addition, any poorly fit real transit signal will not be completely removed through model subtraction, making smaller MES signals very difficult to find. Therefore, masking is the only way to ensure that the previous signal is not being re-detected at each iteration of multiplicity (m). To minimize the effects of masking we choose to remove only $2.5\times$ the transit duration, whereas the *Kepler* pipeline used a mask of $3\times$ the transit duration, which should decrease the probability of masking neighboring transits.

Searching for transits, we limit the search range of the orbital period to $[0.5, 38]$ days. The upper limit of 38 days is set by the span of the data (nominally 74.84 days for Campaign 5), which permits the existence of three transits within the given window. The lower limit was set to minimize contamination from the abundance of harmonic signals that exist at periods <0.5 days (TERRA can detect signals with sufficient MES and periods <0.5 days at a multiple of the true period, but in practice

¹¹ We note that the SES and MES used by TERRA and *Kepler* project Transiting Planet Search (TPS) are analogous, but not strictly equivalent. Differences in the detailed constructions of these statistics may be found by comparing Jenkins et al. (2010) and Petigura et al. (2013).

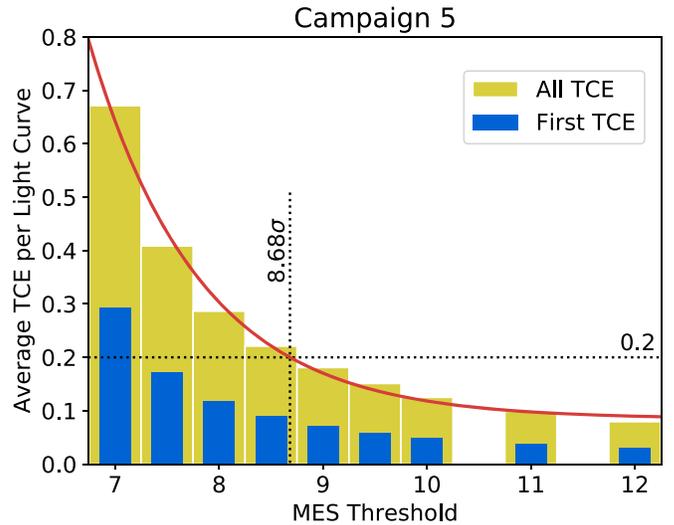


Figure 2. MES threshold plotted against the average number of TCEs found for each light curve in Campaign 5. Many of the light curves will produce more than one TCE, thus the yellow bars (All TCE) represent the total number of detected TCEs divided by the number of light curves searched. The blue bars (First TCE) only consider the first TCE found in each light curve. The red line represents the best fit exponential to the All TCE bars. We select the threshold (8.68) where the average light curve will contain 0.2 TCEs.

we found that these are more often FPs than true planets). Additionally, we wanted to minimize the probability of finding signals artificially introduced by the harmonic fitter (Section 2.2). Finally, of the over 4000 confirmed exoplanets, only 15 have orbital periods <0.5 days (0.37%), including six planets found in the *K2* fields. Thus, limiting our search should have a small effect on the extracted period population. Users concerned with the occurrence of ultrashort-period (USP) planets are encouraged to perform their own custom search.

2.4. Selecting a TCE Limit

The *Kepler* pipeline used a default TCE limit of 7.1σ (Jenkins et al. 2002). Any signal that produces a MES value of this magnitude or greater was considered for vetting. Assuming Gaussian statistics, this limit would only allow one false alarm signal through in the entirety of the *Kepler* mission. However, the distribution of TCEs has long tails, where false alarm detections are more common at 7.1σ than originally expected. *Kepler* found, on average, that 16% of the targets produce a $\geq 7.1\sigma$ event (198,706 light curves; 32,534 TCEs), and a majority of them were later classified as FPs (Thompson et al. 2018).¹²

K2 photometry contains far more outliers than *Kepler* prime photometry, so we tested different MES thresholds to look at the occurrence of TCEs at each limit (Figure 2). The number of TCEs appears to match well with an exponential distribution, and the multiplicity of TCE occurrence does not appear to be affected by a difference in MES threshold. The number of “First TCE” detections, where we only consider the first TCE trigger for each light curve, is roughly 50% of the “All TCE” detections at each of the tested thresholds. This indicates that a light curve where a TCE was triggered was likely to trigger more than one TCE. If this ratio changed significantly at each tested threshold, an argument could be made to select a

¹² In reality *Kepler* detected TCEs in about 8% of the light curves. A large fraction of these light curves contained multiple TCEs, inflating the average.

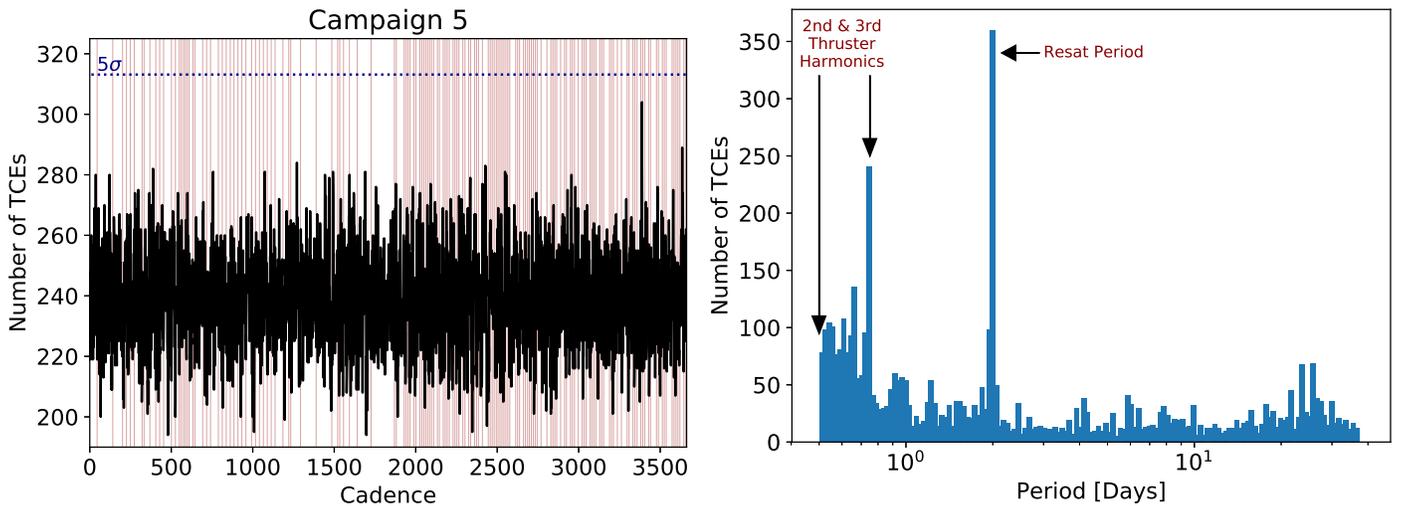


Figure 3. Left: the number of TCEs as a function of cadence. The blue dotted line represents 5σ above the median value. We find no cadences that exceed this limit, and therefore do not apply any additional masking. The red lines indicate the cadences in which a thruster firing occurred. Right: the number of TCEs as a function of period. There are 150 uniform bins in log space. The outlier bins have been labeled to indicate their corresponding cause.

threshold at some level of similarity between the two values. Since no clear convergence occurs, we select a threshold (from our exponential fit) in which on average 20% of the light curves will contain a TCE (8.68σ ; determined by the fit exponential function). Since our threshold is higher than the *Kepler* 7.1σ , we sacrifice some sensitivity in order to minimize the number of false alarm signals (Thompson et al. 2018 found that reliability significantly increased for signals with $MES > 8$ for the *Kepler* pipeline). Previous studies have selected even higher thresholds (9σ ; Kruse et al. 2019), which indicates that our survey will be able to detect smaller and noisier signals at the cost of a higher FP rate.

2.4.1. Skye Excess TCE Identification

One test implemented by Thompson et al. (2018) was the “Skye” metric. Inspired by this metric, we performed a similar analysis, in which we counted up the number of TCEs (N_{TCE}) found at each cadence (cadences considered detected are within the transit duration of the TCE). If all cadences contributed equally to the correlated noise in the light curves, the number of TCEs found at each cadence should be uniform. Any cadences that trigger an abnormally large N_{TCE} were likely faulty and likely to cause FPs, warranting a global removal of this cadence from the time series. Such cadences have been found to be prolific in the *K2* data set, because systematic outliers (due to the pointing jitters) commonly cause false TCE triggering. By manually tuning the threshold across different detrending software, we found that a 5σ threshold provided the best clipping. Thus, we considered a cadence faulty if $N_{\text{TCE}} - \text{med}(N_{\text{TCE}}) > 5\sigma_{N_{\text{TCE}}}$, where $\sigma_{N_{\text{TCE}}}$ is measured using the MAD estimator in Equation (1). We find no cadences that exceed this limit in Campaign 5, but in a preliminary study we found several faulty cadences in the *K2*_{phot} Campaign 5 light curves (Petigura et al. 2015). It appears that the EVEREST software is proficient in removing such cadences. However, we will continue to search for faulty cadences in future campaigns. Figure 3 shows the number of TCEs per cadence, indicating a near-constant distribution.

The numbers of TCEs produced in each period are shown in Figure 3. Two points clearly stand out: 0.75 and 2 days. These two periods correspond to the frequency of various operations

carried out by the spacecraft (Van Cleve et al. 2016). 0.5 and 0.75 days correspond to the second and third harmonics of the thruster firing, which occurs every six hours. Few TCEs are observed at the second harmonic because it sits on the minimum period threshold. Some fraction of the potential TCEs would have been detected below 0.5 days and are now being detected at the next harmonic of 0.75 days. This operation realigned the telescope after drift caused by solar radiation pressure. Although EVEREST masks these data points, the neighboring cadences can cause outliers. Thus, the harmonics of 0.25 days lead to an overabundance of TCEs where these outliers all line up. The jump at two days in Figure 3 corresponds to the frequency of reaction wheel momentum resaturations (known as the “Resat” period). Although many of the cadences were masked by EVEREST, the surrounding cadences can, again, cause a TCE. Our outlier vetting metric is proficient in removing these artificial signals (see Section 3.3)

2.5. Prioritizing Reliability over Completeness

We selected parameters for our pipeline that maximize the completeness and the reliability of our sample. Adopting the philosophy of the *Kepler* DR25 catalog (Thompson et al. 2018), we willingly allow some known planet candidates to achieve an FP label in order to create a uniformly vetted, highly reliable catalog. The parameters described in Section 3 have been tuned to minimize such misclassification, without allowing additional FPs through. We hope to maintain these same parameters across all campaigns, so that an aggregate calculation of occurrence using all 18 campaigns can be made seamlessly.

3. EDI-Vetter

One of the more difficult tasks in any automated transit search is the vetting process. This is the final defense against transit FPs and must be robust against many different types of light-curve anomalies. Automated vetting processes have been created for the *Kepler* pipeline (McCauliff et al. 2015; Thompson et al. 2018), but the *K2* data set is far noisier and additional tests are required. We build upon the criteria used by

the RoboVetter (Thompson et al. 2018) to create EDI-Vetter (Zink 2019), which is more robust to the issues unique to the *K2* data set. A user-friendly version of this software is made publicly available on GitHub.¹³ Here we discuss the metrics EDI-Vetter uses to ensure our candidate list is pure with high reliability.

To begin, our software uses the maximum likelihood parameters found by TERRA to fit a transit model to the detected TCE signal. Using the affine invariant sampler of Goodman & Weare (2010), as implemented in Python by Foreman-Mackey et al. (2013) (emcee), and the batman transit model (Mandel & Agol 2002; Kreidberg 2015), we confirm that the TERRA values found a global maximum likelihood and determine the maximum likelihood of additional transit parameters. We produce our transit model with the batman Python package (Kreidberg 2015). Using 100 semi-independent walkers, 250 burn-in steps, and 100 parameter test steps (35,000 total steps; 10,000 test samples), we estimate the maximum likelihood and uncertainty for the transit ephemeris (t_0), the radius ratio (R_{p1}/R_*), the impact parameter (b), the signal period (P), the ratio of semimajor axis to stellar radius (a_{p1}), the transit duration (t_{dur}), and the transit depth. We assume a circular orbit and derive two quadratic limb darkening coefficients using the stellar parameters available on ExoFOP¹⁴ (Huber et al. 2016) along with the ATLAS model table for *Kepler* bandpass limb darkening coefficients from Claret & Bloemen (2011). According to the literature for emcee,¹⁵ 50 times the integrated autocorrelation time (τ) is the lowest limit at which the Markov Chain Monte Carlo (MCMC) sampler results can be trusted. We find $\tau \sim 50$ samples for our five-parameter model, indicating that we should require at least 2500 samples before our results can be considered meaningful. Clearly, the 10,000 samples taken by our MCMC sampler is near this lower limit and likely insufficient for a thorough parameter estimation, but we sacrifice some accuracy to significantly improve the computational speed of our software. For vetted candidates we run a more thorough parameter estimation as described in Section 8.

In the next 13 Sections 3.1–3.13 we provide a detailed description of the filters used by EDI-Vetter as shown in Figure 4.

3.1. Previous Planet Check

This test ensures that the detected TCE is not a repeat of a previous signal (Section 3.2.2 of Coughlin et al. 2016). TCE recurrence is common for FPs, where masking is inefficient at removing the entirety of the light-curve anomaly. Additionally, true signals that have periods misidentified by some integer multiple will produce multiple TCE signals. This test helps remove such redundancies. If the pipeline finds more than one TCE for a given light curve, this test will be enacted in decreasing signal detection order, i.e., the first signal detected will not be subject to this test, the second signal will be tested against the first, and the third signal will be tested against the first and the second signals.

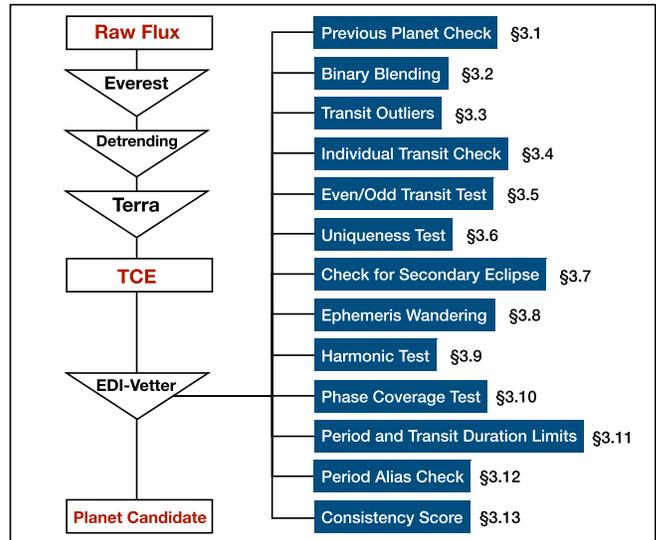


Figure 4. A schematic of our *K2* pipeline. The section numbers correspond to the detailed description of the EDI-Vetter metric.

We define ΔP as the separation in periods, normalized by the shorter of the two periods:

$$\Delta P = \frac{P_B - P_A}{P_A} \quad (2)$$

where P_A and P_B represent the shorter and longer periods respectively. To then determine the offset from the nearest integer multiple, δP is calculated:

$$\delta P = |\Delta P - \text{round}(\Delta P)| \quad (3)$$

where the round() function will round to the nearest integer value. We can then determine how statistically significant (σ_P) a period separation is using the erfcinv() (inverse complementary error function):

$$\sigma_P = \sqrt{2} \text{erfcinv}(\delta P) \quad (4)$$

where larger δP values will produce small σ_P values. For consideration as a candidate we require signals with a high detection order (signals found after the first signal within the light curve) to be separated with $\sigma_P \leq 2$ for candidate consideration. If a signal matches with a previous detection with $\sigma_P > 2$, it is likely a repeated detection. However, we also must consider the case of resonant systems, where integer period separations are common. To avoid falsely eliminating such systems, we also consider the ephemeris separation for the two signals (Δt_0):

$$\Delta t_0 = \frac{|t_{0A} - t_{0B}|}{t_{dur}} \quad (5)$$

where t_{dur} is the transit duration of the signal in question and t_{0A} and t_{0B} are the ephemeris times for the first transit within the data set. If Δt_0 is large (≥ 1) and ΔP is small (≤ 2), it is possible that the signals are from a resonant orbit. The caveat to this is the case where the signal with the higher detection order is fitting to the secondary eclipse of the transit. To ensure this is

¹³ <https://github.com/jonzink/EDI-Vetter/>

¹⁴ <https://exofop.ipac.caltech.edu>

¹⁵ <https://emcee.readthedocs.io/en/latest/tutorials/autocorr/>

not the case ΔSE_1 and ΔSE_2 are calculated:

$$\begin{aligned}\Delta SE_1 &= \frac{|t_{0A} - t_{0B} + P_A/2|}{t_{\text{dur}}} \\ \Delta SE_2 &= \frac{|t_{0A} - t_{0B} - P_A/2|}{t_{\text{dur}}}\end{aligned}\quad (6)$$

Again, a small ΔSE_1 or ΔSE_2 (<1) indicates that the higher order detection is likely fitting to the secondary eclipse. This metric will only be sensitive to eclipsing binaries with a circular orbit, where the secondary eclipse is at a phase of 0.5. Future iterations of this pipeline will consider non-circular orbits in this metric.

All of these metrics are assuming that the previous planet was a real signal. If the previous detection was deemed an FP and $\sigma_P > 2$, this signal is likely just a repeated detection from an undermasked anomaly. Thus, the signal will be flagged as an FP.

For clarity, we summarize here the full process in which a signal will be classified as an FP. If $\sigma_P > 2$ and the previous signal was deemed an FP, this detection will be flagged as an FP. If $\sigma_P > 2$ and $\Delta t_0 < 1$, this detection will be flagged as an FP. If $\sigma_P > 2$ and $\Delta SE_1 < 1$ or $\Delta SE_2 < 1$, this detection will be flagged as an FP. Otherwise, this signal will continue to be considered for candidacy.

3.2. Binary Blending

In this metric we seek to identify cases where the signal is due to an eclipsing binary, either from the target or from a nearby source, but with such a large impact parameter and/or dilution from the third light source that the observed depth is comparable to that of a transiting planet. During the original *Kepler* mission, flux contamination from nearby sources was ruled out by fitting the pixel response function model (Bryson et al. 2010) to the star in and out of transit, looking for offsets in the photocenter of the light (the centroid). If statistically significant differences appeared, the target was considered contaminated (Mullally 2017). Two scenarios could result in such a shift. First, contamination from a source within a few arcseconds of the primary target would produce this type of shift during transit. Second, a very deep transit signal from a nearby eclipsing binary could contaminate a static light curve, causing a shift during transit. Such methods are difficult to apply to *K2* because of the roll experienced by the telescope, which moves the target image across several pixels. Using cadences with similar roll angles, the DAVE software is able to recoup some of this detection ability (Kostov et al. 2019). However, the shorter data span of each *K2* campaign, and the lack of similar data on roll angle, makes finding these statistical differences more difficult. Instead, we use the *Gaia* DR2 data set (Gaia Collaboration et al. 2018) to look for flux contamination from nearby sources.

Gaia can resolve nearby sources down to $1''$ for *G*-band $\Delta \text{mag} \lesssim 3$ (Ziegler et al. 2018). Since the *K2* pixel scale is $3''.98$, we can achieve sub-pixel resolution using cross-matching alone. We utilize the `gaia-kepler.fun`¹⁶ cross-match database with a $20''$ search radius to look for these contaminants. To ensure we have correctly matched the target star, we require the “`phot_g_mean_mag`” parameter to be within 1 magnitude of the assigned “`k2_kepmag`.” These

values are often almost indistinguishable because the *Gaia* *G* band and *Kepler* band probe nearly the same wavelength range. If this criterion is not met, we check the next closest source within the $20''$ search radius to see if it meets this criterion. This process will continue until a target of the correct magnitude is found or all of the sources in the $20''$ search radius have been tested. If no source is within this 1 magnitude limit, we will select the source that provides the closest magnitude match. This closest-match scenario was applied to 301 of the C5 targets (1.2% of the total tested targets).

As used for the EVEREST pipeline (Luger et al. 2018), we adopted aperture #15 from the *K2SFF* catalog (Vanderburg & Johnson 2014). Using the size of the aperture, we can determine the amount of flux contamination from neighboring binaries. Often the apertures select pixels in a circular manner around the target, but in cases where additional point sources are nearby, the aperture may be elongated, capturing additional flux from the nearby source. To account for this possible elongation, we consider flux contamination within this radius (Δap):

$$\Delta \text{ap} = 3''.98 \sqrt{\frac{N_{\text{pix}}}{\pi} + 1} \quad (7)$$

where N_{pix} is the number of pixels for the given aperture. This Δap is likely overestimating the flux contamination for many of the well-behaved apertures, where the aperture is large and circular, but it is important in order to ensure high reliability. To determine the amount of flux contamination, we compute the flux ratio for the nearest potential contaminating source ($F_{\text{Gaia}}^{\text{Ratio}}$):

$$\frac{F_{\text{neighbor}}}{F_{\star}} = F_{\text{Gaia}}^{\text{Ratio}} = 10^{\frac{\Delta m_{\text{Gaia}}}{-2.5}} \quad (8)$$

where F_{neighbor} is the expected flux from the nearby source, F_{\star} is the expected flux from the target star, and Δm_{Gaia} is the difference in magnitude between the target source and the possible contaminant in the *Gaia* *G* band. To determine the impact of this external source on our target, we calculate the ratio of total flux to the target flux ($F_{\text{Gaia}}^{\text{Tot}}$):

$$F_{\text{Gaia}}^{\text{Tot}} = \frac{F_{\star} + F_{\text{cont}}}{F_{\star}} = 1 + \frac{F_{\text{Gaia}}^{\text{Ratio}}}{2} \left(1 + \text{erf} \left(\frac{\Delta \text{ap} - \Delta d}{2''.55\sqrt{2}} \right) \right) \quad (9)$$

where F_{cont} is the fraction of flux from F_{neighbor} within the aperture of the target and Δd represents the angular distance between the target and the contaminating source. Using the error function in this method assumes a 1D Gaussian point-spread function (PSF) with a standard deviation of $2''.55$ (corresponding to the $6''$ FWHM of the *Kepler* PSF). The 1D assumption will certainly overestimate the amount of flux contamination, but this cautious approach will improve our reliability against astrophysical FPs. For targets with multiple neighboring stars we consider the flux contamination from all the sources within a $20''$ radius of the source.

One limitation to this approach is that our search limit is bound by the $20''$ search radius of the `gaia-kepler.fun` cross-match. In some rare instances, an aperture will extend beyond this radius ($N_{\text{pix}} > 78$). Upon visual inspection, most of these targets are in very crowded fields with several stars within a single aperture. Implementing this aperture limit, we remove 54 targets from our search. Additionally, contaminating sources

¹⁶ <https://gaia-kepler.fun>

can also exist beyond the $20''$ search radius. To help account for these cases, we also use the J -band photometry of the Two Micron All Sky Survey (2MASS). Pulling from ExoFOP (see footnote 14), which provides 2MASS cross-matches beyond $20''$, we calculate Equation (8) for 2MASS and find the potential contamination (F_{2MASS}^{Tot}):

$$F_{2MASS}^{\text{Tot}} = 1 + \frac{F_{2MASS}^{\text{Ratio}}}{4} \left(1 + \operatorname{erf} \left(\frac{\Delta \text{ap} - \Delta d}{2'' 55 \sqrt{2}} \right) \right). \quad (10)$$

Here, it is important to note that we assume the flux contamination will be half as strong as a similar contaminant in the *Gaia* band. This is to account for the fact the 2MASS J -band photometry cannot be directly compared to the expected $K2$ flux. Howard et al. (2012) suggests a flux ratio of $\sim 1/4$ for a J -band to K -band conversion, but we choose $1/2$ to err on the side of overestimating potential contamination. We find that F_{2MASS}^{Tot} is usually very close to unity; only in extreme cases where the J -band contamination is significant, and $F_{Gaia}^{\text{Tot}} = 1$, will the 2MASS flux ratio play a role.

To determine the impact of this flux contamination on our source, we consider how F_{Gaia}^{Tot} and F_{2MASS}^{Tot} affect the inferred planet radius. One of the biggest concern with binary contamination is the potential to decrease the transit depth and falsely present an eclipsing binary as a planet candidate. Additionally, in systems with a transiting planet, the planet radius is derived from the depth of the measured transit. When flux from an additional star is present, the depth of the transit will be diluted and result in an underestimation of the planet radius (Ciardi et al. 2017; Fulton et al. 2017; Matson et al. 2018). We adopt the same requirements on radius and impact parameter as suggested by Batalha et al. (2013), but now include a correction for flux contamination:

$$\frac{R_{\text{pl}}}{R_{\star}} \sqrt{\max(F_{Gaia}^{\text{Tot}}, F_{2MASS}^{\text{Tot}})} + b \leq 1.04 \quad (11)$$

where b is the impact parameter and R_{pl}/R_{\star} is the radius ratio fit to the transit in question. The max function chooses the higher of the two flux ratios. This ensures that we do not double-count the flux contamination from a source. The median F^{Tot} value for targets in Campaign 5 is 1.00003, indicating that most targets are well isolated. Any signal that exceeds the provided limit (Equation (11)) will be flagged as an FP. This method of FP flagging allows us to remain agnostic to the stellar parameters. Additionally, we put an upper limit on R_{pl}/R_{\star} . Any TCE found with $R_{\text{pl}}/R_{\star} > 0.3$ will automatically be flagged as an FP. This limit is imposed to eliminate eclipsing binaries with low impact parameter.

3.3. Transit Outliers

One common issue seen in many of the TCEs found in the $K2$ light curves is the chance alignment of several outliers. This metric is aimed to identify when such cases occur. The general increase in noise caused by the spacecraft roll motion and thruster firing caused a significant number of outliers to occur in the flux data set. Many of these are removed by the EVEREST processing, but occasionally a few will make it through (evidence of this can be seen in the right panel of Figure 3). When this occurs, the TCE search is likely to find a period where the outliers line up when phase-folded. These

chance alignment signals are hard to distinguish with previously developed vetting tools, but easy to spot by eye. Here we present a unique test to eliminate these anomalies.

We use the best fit model to subtract the signal from the folded data set. We first check to see whether the in-transit residuals have a larger variance (σ_{Tran}) than the rest of the light curve (σ_{Base}). However, this calculation is complicated by larger MES signals. We found that many high-MES transit signals will produce larger variances in the residuals because of either poor fitting or poor detrending. Fortunately, many of these chance alignment outliers produce a low MES signal. Thus, we deem any transit residual with $\sigma_{\text{Tran}} > (0.461 \times \text{MES} - 2)\sigma_{\text{Base}}$ an FP. This threshold was tuned in correspondence to an acceptable increase of $2\times$ the in-transit rms at the detection limit and $4\times$ the rms at $\text{MES} = 13$. In Figure 5 we show two signals—one that passed this test and one that failed.

We often find that several aligned outliers will trigger a TCE, but retain a low σ_{Tran} due to a larger number of non-outlier baseline flux measurements included in the σ_{Tran} calculation. To account for these cases, we count up the number of in-transit cadences that produce outliers (residual $> 3\sigma_{\text{Base}}$) in the transit residuals. If this number (Out) exceeds the following limits, the signal is deemed an FP:

$$\begin{aligned} \text{if } \text{MES} < 40: & \text{Out} > N_{\text{Tran}} \\ \text{else:} & \text{Out} > 0.3 \times \text{MES} + 4 \end{aligned} \quad (12)$$

where N_{Tran} is the number of transits. Equation (12) allows, on average, one outlier per transit for signals with $\text{MES} < 40$. For stronger signals, outliers are more likely to be triggered due to poor model fitting. This problem becomes more significant as MES increases, thus we use the linear function of MES as a threshold for these cases. This function was tuned to allow 16 outliers at $\text{MES} = 40$ and 25 outliers at $\text{MES} = 70$. This metric performs well against outlier alignments, but has a mild tendency of flagging high-MES, but poorly detrended, planets as FP. This is the unfortunate cost of attaining a high-reliability catalog.

3.4. Individual Transit Check

Much of this part of our pipeline is derived from the ‘‘Marshall’’ test (Mullally et al. 2016). For clarity we will briefly describe how this test works. We look at each transit within the light curve individually. Each transit detected is then fit to four potential models:

$$\begin{aligned} f(t) &= y_0 && \text{Flat} \\ f(t) &= y_0 + S(t) && \text{Logistic} \\ f(t) &= y_0 + S(t)(1 - e^{-\beta t}) && \text{Logistic-exponential} \\ f(t) &= y_0 + S(t - \tau/2) - S(t + \tau/2) && \text{Double-logistic} \end{aligned} \quad (13)$$

where y_0 is a constant offset and τ and β are tunable parameters (see Figure 1 of Mullally et al. 2016 for reference). We also do not fit an additional GP (as noted in the original Marshall test) because the light curves have already experienced several GP fitters. The function $S(t)$ is a logistic function given as

$$S(t) = \frac{d}{1 + e^{\gamma(t-t_0)}} \quad (14)$$

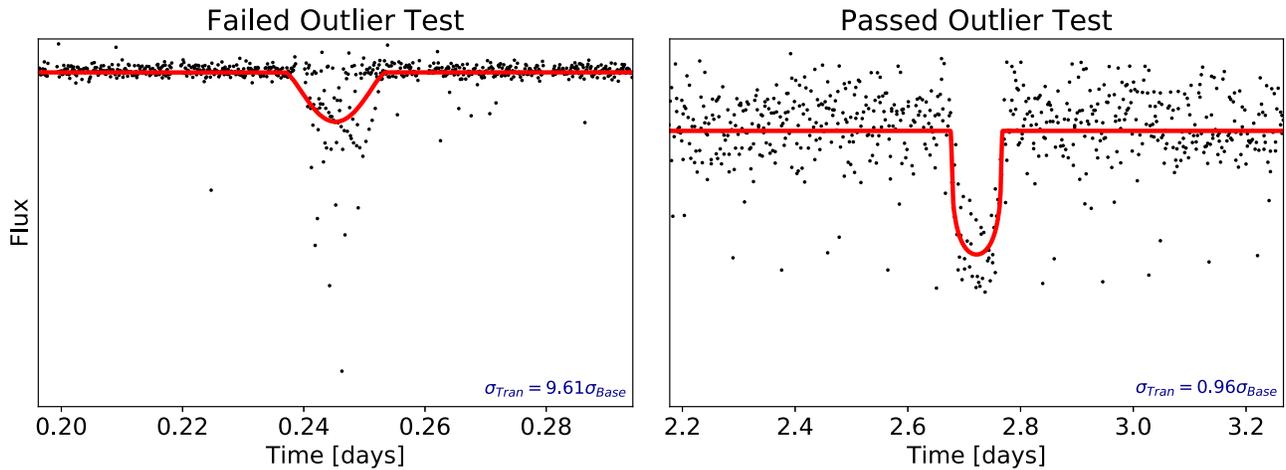


Figure 5. Left: a TCE with a period of 0.5 days, likely caused by the alignment of measurements made near thruster firings. The large in-transit rms (σ_{Tran}) is an indicator that the signal is an FP. This TCE was rejected using the outlier variance test. Right: EPIC 211331236.01 with a period of 5.4 days. The in-transit deviations match well with the out-of-transit fluctuations. Thus, this signal was not removed by our outlier variance test.

where t_0 , γ , and d are all tunable parameters. These functions are meant to replicate possible FP signals. To gauge the success of each fit, we calculate the Bayesian information criterion (BIC) for each function. This value is meant to mimic the Bayesian evidence of a Gaussian likelihood, and penalizes for increasing the number of tunable parameters. We also calculate the BIC for our best fit transit model. If the BIC value for any of the previously noted functions is 10 less than the transit BIC ($\text{BIC}_{f(t)} + 10 < \text{BIC}_{\text{Transit}}$) and $\text{MES}/\sqrt{N_{\text{Tran}}} > 4$ (this criterion ensures that the expected single transit signal strength is large enough to provide a meaningful model fit), we mask that transit signal. However, the requirement of $\text{MES}/\sqrt{N_{\text{Tran}}} > 4$ is ignored for the model with constant y_0 .

We also consider the number of unmasked points in each transit. If more than 50% of cadences within the transit duration are masked, the entire transit is masked. This ensures that a significant fraction of the transit data is available. Here we treat all parts of the transit signal equally, but in reality the egress and ingress are more important for candidacy. We consider this further in Section 3.10. The current test has the potential to mask several transits. To see how this affects the signal we reanalyze the light curve to see whether the transit still has three observable transits and $\text{MES} \geq 8.68$. If the signal now fails to meet either of these requirements, it is flagged as an FP.

Looking at the single event statistics (SES) of each transit, we can determine whether a single transit is dominating the MES. In an ideal case $\text{SES} = \text{MES}/\sqrt{N_{\text{Tran}}}$. EDI-Vetter will look for SES values that exceed 80% of the overall MES value. In the scenario where only three transits exist we would not expect any transit to exceed $\sim 60\%$, thus 80% provides an unlikely natural occurrence. As performed in Section 3.2.4 of Coughlin et al. (2016) and Section A.3.5 of Thompson et al. (2018), this test will help eliminate single transit outliers that pushed the MES over the TCE limit. It is not robust to chance alignments of multiple transit outliers as addressed in Section 3.3. If this limit ($\text{SES} > 0.8\text{MES}$) is exceeded, the signal is deemed an FP by the pipeline.

One potential issue with this test is that a true planet transit that falls on a systematic feature could inflate the SES beyond this threshold, triggering an FP flag. In testing, this was not an apparent issue, but it is something we will continue to monitor as we move forward with the other campaigns.

3.5. Even/Odd Transit Test

This classic test will look for cases where the transit primary and secondary eclipses are folded on top of each other. This will occur when the secondary eclipse depth is a significant fraction of the primary one and the orbit is circular so that the secondary eclipse occurs at a phase of 0.5. We would only expect such similarity between transit depths for eclipsing binaries. Thus, when we find overfolded light curves of this nature, we flag them as FPs.

To perform this test, we separate the phase-foldings into even and odd transits (i.e., every other transit in one group and the remainder in the other group). The transits are then refit, holding P constant. We compare the inferred values of R (radius ratio) from each group (odd versus even) to look for statistical differences (Z):

$$Z = \frac{R_{\text{even}} - R_{\text{odd}}}{\sqrt{\sigma_{R_{\text{even}}}^2 - \sigma_{R_{\text{odd}}}^2}} \quad (15)$$

where the σ values correspond to the labeled R fit uncertainties. We select $Z > 5$ for FP flagging.

One notable issue arises when the true period of the signal is twice the period detected, folding baseline data on top of the transit signal. This will be flagged as FP despite being a true planet signal. Cases where this occurs will be identified in Section 3.12 and rerun at the correct period folding.

One of the main complications of this test are the cases where one of the transits (usually near the end or beginning of the campaign) is poorly detrended. If this occurs in a true signal with very few transits ($N_{\text{Tran}} \leq 5$), one of the two groups can be significantly pulled by this faulty transit and cause this flag to be falsely triggered. Many of the known high-MES candidates not found by our pipeline fall victim to this flagging. This problem can be easily spotted by eye, but remains difficult to remove automatically.

3.6. Uniqueness Test

This metric identifies cases where the phase-folded light curve appears to produce several transit-like dips. This type of noisy light curve is unlikely to reveal a real planet signal, but rather highlights the systematic issues of the spacecraft. We base this metric on the “model-shift uniqueness test”

(Mullally et al. 2015; Rowe et al. 2015; Coughlin et al. 2016; Thompson et al. 2018). Here, we provide a brief explanation of this test. For a thorough explanation, we refer interested readers to Coughlin (2017).

The first step in this process is to fold the light curve according to the period of the signal in question. The light curve should remain folded on this period for the entirety of this test. To determine whether a signal is unique to the light curve, it is first important to mask the detected signal from the light curve and any possible secondary eclipse. This is done by masking the largest signal (MES_1) and the second largest signal (MES_2) in the light curve, given that the second largest signal is within 50% of the transit duration (t_{dur}) of the expected secondary eclipse location. If the signal MES_2 is not within the expected secondary eclipse window, the third largest signal (with the same t_{dur}) is assigned to MES_2 before masking. This allows the second largest signal to remain in the light curve for detection as MES_3 . We can then measure the red noise within the period-folded data. This is achieved by fitting a transit model to each flux point with a fixed width of the original signal duration (t_{dur}). We simplify this process by using a box-shaped transit, where the best fit depth can be analytically shown to be the mean of the selected points. This simplification significantly increases the speed of our processing. If the light curve has no red noise, these averaged depth values will have a standard deviation (σ_{Red}) of the order of the light curve noise (σ_{LC}); otherwise it will be much larger. We then calculate the ratio of these two values (F_{Red}):

$$F_{Red} = \sigma_{Red} / \sigma_{LC}. \quad (16)$$

Additionally, we want to determine the strength of the third largest signal (MES_3) within the light curve, and then the largest flux brightening signal (the largest signal found when the light curve is inverted: MES_4).

For statistical comparison, we consider the threshold in which a signal is statistically significant under the assumption of Gaussian noise:

$$F_1 = \sqrt{2} \operatorname{erfcinv}\left(\frac{t_{dur}}{P \times N_{TCE}}\right) \quad (17)$$

where P is the period in which the light curve has been folded, and N_{TCE} is the number of TCEs expected in the entire campaign. In an effort to ensure high reliability, we select $N_{TCE} = 25,000$, or roughly one TCE per light curve. Here P/t_{dur} represents the number of events in each light curve and N_{TCE} is the number of light curves considered. For the statistical significance within a single light curve, we consider the threshold as follows:

$$F_2 = \sqrt{2} \operatorname{erfcinv}\left(\frac{t_{dur}}{P}\right). \quad (18)$$

We consider a signal viable for further vetting if all the following inequalities are satisfied:

$$\begin{aligned} F_1 - (MES_1 / F_{Red}) &\leq 1 \\ F_2 - (MES_1 - MES_3) &\leq 2 \\ F_2 - (MES_1 - MES_4) &\leq 3. \end{aligned} \quad (19)$$

If any of the previous equations is false, the signal is flagged as an FP (see Figure 19 of Coughlin 2017 for reference). The inequality threshold values are based on those provided in

Thompson et al. (2018), but have been manually tuned to suit the needs of this pipeline.

We also consider the significance of the secondary eclipse detected within the folded light curve. If the phase of the secondary eclipse falls within $0.5t_{dur}$ of $t_0 + P/2$, and $MES_1 - MES_2 - F_2 > 0$, the secondary eclipse appears genuine and warrants further testing in Section 3.7. This signal will be flagged with a secondary eclipse flag. It is important to note that this test will only be sensitive to companions with near-circular orbits, where the secondary eclipse occurs at or near a phase of 0.5. Future iterations of the pipeline will consider eccentric orbits. If the secondary eclipse does not fall at or near a phase of 0.5, the signal is deemed an FP due to lack of uniqueness if all of the following inequalities are satisfied:

$$\begin{aligned} MES_2 / F_{Red} - F_1 &> 1 \\ MES_2 - MES_3 - F_2 &> 0 \\ MES_2 - MES_4 - F_2 &> 0. \end{aligned} \quad (20)$$

If any of the previous equations is false, the signal will be further vetted for candidacy. If the secondary eclipse is not astrophysical, the inequalities in Equation (20) make certain that the original MES_1 signal is still sufficiently unique.

3.7. Check for Secondary Eclipse

Apparent secondary eclipses are often the hallmark of an eclipsing binary. However, such events can also be seen for certain transiting hot-Jupiter-like planets. In the previous section, we discussed the criteria for a secondary eclipse to be considered statistically significant. If the TCE was flagged as a secondary eclipse candidate, we consider several metrics before determining how to classify the signal.

Inspired by the framework of Section A.3.2 in Thompson et al. (2018), we begin by fitting a transit model to the secondary eclipse in question. If the depth of the secondary eclipse in that model is greater than 10% of the depth of the initial transit, the radius ratio of the secondary eclipse is statistically significant ($R_{SE} > 2\sigma_{R_{SE}}$), and the initial impact parameter of the transit $b \geq 0.8$, then the TCE is flagged as an FP. The large secondary eclipse and high impact parameter are indicators that the transit is likely a grazing eclipsing binary. If the signal fails to meet any of the listed criteria, the TCE remains labeled as a secondary eclipse candidate and continues for further vetting. Such cases are likely transiting hot-Jupiters reflecting a significant amount of light. It is possible that a well aligned eclipsing binary with a low impact parameter could potentially slip through this metric, but the occurrence of such an event would likely produce a large enough transit depth to trigger the metrics discussed in Section 3.2.

3.8. Ephemeris Wandering

The ephemeris value (t_0) found by the grid TCE search (TERRA) will be very similar to the value determined by the transit model fitting if the signal is a true planet candidate. In cases where the two values differ significantly, the signal is likely very asymmetric and does not fit well with the transit model. To avoid these contaminants, we flag all signals as FP if $|\Delta t_0|$ between the grid search and the model fitting is greater than $0.5t_{dur}$.

3.9. Harmonic Test

Sinusoidal stellar variability is a common trigger for TCE search algorithms. In Section 2.2 we attempted to detrend such periodicity, but our limits of $P < 0.5$ days for the harmonic fitter and $P > 5$ days for the GP fitter create a window in which a sinusoidal signal can sneak in. Furthermore, cases where the sinusoidal signal is not completely removed can cause TCE triggering. To eliminate such contamination, we fit several sine curves to test for such FPs.

$$\eta(t_{\text{mid}}) = \begin{cases} t_{\text{cad}} \times (16\phi(t_{\text{mid}})^4 - 8\phi(t_{\text{mid}})^2 + 2); & \text{if } -1 \leq \phi(t_{\text{mid}}) \leq 1, \\ \infty; & \text{else} \end{cases} \quad (22)$$

When testing the possible harmonic functions, we fit a sine curve to the data while holding fixed the period and allowing the phase and amplitude (A_{sin}) to vary. We attempt to fit six different harmonic periods: P (the TCE period), $P/2$, $2P$, t_{dur} (the transit duration), $2t_{\text{dur}}$, and $4t_{\text{dur}}$. These potential periods represent the most notorious harmonics found mimicking a TCE.

When considering how well the sine curves fit the data, it is important to only consider the A_{sin} value. Any attempt to use the BIC value, as done in Section 3.4, can cause one to misclassify hot-Jupiters, which can naturally produce strong sinusoidal oscillations within the folded light curve. If any of the tested periods produce $A_{\text{sin}}/\sigma_{A_{\text{sin}}} > 50$, the signal is labeled an FP. In cases where the harmonic signal is caused by a hot-Jupiter, the primary and secondary eclipses will reduce the sine signal strength below 50σ . Furthermore, a large amplitude in itself may be an indicator that the oscillations are of stellar origin. Our second test checks to see whether any of the A_{sin} values are larger than 90% of the fit transit depth and $A_{\text{sin}}/\sigma_{A_{\text{sin}}} > 2$, ensuring the signal is real. If this criterion is met, the signal is classified as an FP. Large A_{sin} values correspond to a harmonic that is contributing to a large fraction of the signal depth and is likely the cause of the TCE.

As mentioned in Section 2.2, part of our preprocessing attempts to fit harmonic trends in the light curve. To account for the possibility of artificially introducing a signal that could be detected as a TCE, we calculate the power (Pow_{TCE}) of a Lomb–Scargle periodogram at the transit period (P), and then search for the largest signal (Pow_{max}) with a period less than the transit period but greater than twice the cadence spacing. The goal here is to determine whether the TCE is some integer multiplier of the initial harmonic fitter. If $\text{Pow}_{\text{max}} > (\text{Pow}_{\text{TCE}} + 0.5)$ and the Pow_{max} signal is found with a period less than 0.5 days, the TCE is deemed an FP.

3.10. Phase Coverage Test

With the significant processing required to detrend the *K2* light curves, on average $\sim 2\%$ of the data are masked out on any given light curve due to systematic issues causing flux outliers. Although the test performed in Section 3.4 will remove transits that have less than 50% phase coverage, it is still possible for phase-folded data to contain large gaps in the transit signal. This occurs when all the masked portions of the light curve line up in phase space. This is especially common for planets with periods that are near integer multipliers of the *K2* cadence. It is difficult to properly vet such cases, because the transit appears incomplete

and prone to FP contamination. Thus, we have developed a metric for flagging TCEs with such gaps.

Chance outlier alignments lack an important trait: they do not have a pronounced ingress and egress. To capitalize on this feature, we require more phase coverage during these periods than during the transit midpoint. The phase-folded gap allowance ($\eta(t_{\text{mid}})$) is as follows:

$$\phi(t_{\text{mid}}) = \frac{(t_{\text{mid}} - t_0) - \text{round}(t_{\text{mid}} - t_0)}{t_{\text{dur}}} \quad (21)$$

where t_{mid} is the midpoint between the nearest-neighbor measurements in the phase-folded time series, t_0 is the ephemeris time, t_{dur} is the duration of the fit transit, and t_{cad} is the cadence of the light curve (nominally 0.0204 days). $\eta(t)$ provides an allowance of t_{dur} during ingress and egress and a value of $2t_{\text{cad}}$ at the ephemeris. $\eta(t_{\text{mid}})$ quickly becomes very large for t_{mid} values beyond t_{dur} ($\eta(\phi = 1, -1) = 10t_{\text{cad}}$), but still requires some baseline coverage immediately before ingress and after egress. If $\Delta t_{\text{mid}} > \eta(t_{\text{mid}})$ (where Δt_{mid} is the temporal spacing between the two measurements that define t_{mid}) for any value of t_{mid} , the TCE is flagged as an FP. This metric will flag TCEs with large gaps in the phase-folded transit signal. As mentioned in Section 2.4.1, the EVEREST software masks many of the cadences of spacecraft operation, but the neighboring cadences can occasionally produce outliers that line up to form TCEs. Such cases will produce gaps in the phase-folded transit signal and be flagged as an FP by this metric.

Additionally, to ensure large gaps do not exist in the transit phase coverage, similar to what was done in Section A.3.7.1 of Thompson et al. (2018), we implement a requirement of a minimum 70% in transit phase coverage. Any TCE that contains phase-folded temporal gaps larger than $0.3t_{\text{dur}}$ will be flagged as an FP.

3.11. Limits on Period and Transit Duration

As mentioned in Section 2.2, we used a harmonic fitter to remove stellar noise at periods less than 0.5 days. To avoid contamination caused by the introduction of this harmonic, we limit our sample to $P \geq 0.5$ days. This limit has been enforced on TERRA in Section 2.3, but such grid search algorithms can still provide detections beyond this limit by slightly misidentifying the true transit period. Furthermore, TCEs found with periods very near 0.5 days can move below this limit when implementing the parameter search method of EDI-Vetter. We strictly enforce this limit by again checking that $P \geq 0.5$ days; otherwise the TCE is deemed an FP.

Another common FP is when a stellar oscillation has a short period (< 1 day) falling in the forbidden detrending window (noted in Section 2.2). In addition to our harmonic test (Section 2.2), a symptom of a sine-wave TCE is a very long t_{dur} when compared to the signal period. Thus any TCEs with $t_{\text{dur}}/P > 0.1$ are flagged as FPs. This also ensures that the planetary parameters remain physical. Any transit with

$t_{\text{dur}}/P > 0.1$ will be orbiting very near the surface of the star ($a_{\text{pl}} < 3 R_{\star}$).

We acknowledge that ultrashort-period planets (Sanchis-Ojeda et al. 2014; Adams et al. 2016) will certainly be rejected by the discussed cuts. However, all attempts made to capture these missed planets introduced a significant number of FPs into our candidate sample, negating our goal of high reliability. Again, readers concerned with USP planets should perform their own fine-tuned searches.

3.12. Period Alias Check

Occasionally the TERRA grid search will find a transit that is some integer multiple of the true signal period. In such cases, it is very common that the signal is deemed an FP because of the significant “secondary eclipse,” or the transit is poorly fit, causing several other FP flags to be triggered. Alternatively, a signal with a period less than 0.5 days can be found with some larger multiple of the true period and can contaminate the candidate population.

To avoid such cases, we measure the Gaussian likelihood ($\log \mathcal{L} = -\frac{1}{2}\chi^2$) of the transit model with the fit transit period (P), and compare that with the likelihood of the same model with four other possible periods: $P/2$, $P/3$, $2P$, and $3P$. We also scale the ratio of semimajor axis to stellar radius (a_{pl}) accordingly to ensure the transit duration is consistent among all tested likelihoods. To avoid edge cases where noisy data could falsely score one of the likelihoods higher, we give the initially fit period a slight advantage. If any of the alternative periods produces a likelihood greater than 1.05, the likelihood of the initially fit transit period, the signal is flagged as a period alias.

Any signal flagged as a period alias will be refit and revetted with the corresponding highest likelihood period. This cycle will occur at most three times (in most cases it only occurred once), allowing for higher-order integer multipliers to be tested. Within each cycle, the FP flag will be reset and the vetting metrics will be reprocessed at the new period. We then take the vetting as it stands in the final cycle.

3.13. Consistency Score

One cannot help but worry about the edge cases that could potentially slip through the cracks of our vetting metrics. To combat such cases, we implement a limit on consistency score. To calculate the consistency score of our vetted candidates, we run each of them through the EDI-Vetter fit and vetting pipeline 20 times. Given the statistical nature of the MCMC parameter estimation (and the limited number of burn-in and test steps), each run will process the light curve slightly differently, potentially pushing the candidate into the FP bin. We only label a TCE as a candidate if it comes out of the vetting metric as an FP 25% of the time or less (consistency score ≥ 0.75). This metric is comparable to the “disposition score” made available for *Kepler* DR25 (Thompson et al. 2018).

4. CDPP

In Christiansen et al. (2012) the combined differential photometric precision (CDPP) metric was defined to create a quantitative measure of the expected stellar variability and systematic noise. In other words, the CDPP tells us how strong a signal must be in order to overcome the noise present in the

Table 1
Description of the CDPP and Measurements of Window Function for Each Stellar Target

Column	Format	Units	Explanation
1	I9	...	EPIC identifier
2	F11.3	ppm	CDPP rms value for transit of 1.0 hr
3	F11.3	ppm	CDPP rms value for transit of 1.5 hr
4	F11.3	ppm	CDPP rms value for transit of 2.0 hr
5	F11.3	ppm	CDPP rms value for transit of 2.5 hr
6	F11.3	ppm	CDPP rms value for transit of 3.0 hr
7	F11.3	ppm	CDPP rms value for transit of 4.0 hr
8	F11.3	ppm	CDPP rms value for transit of 5.0 hr
9	F11.3	ppm	CDPP rms value for transit of 6.0 hr
10	F11.3	ppm	CDPP rms value for transit of 7.0 hr
11	F11.3	ppm	CDPP rms value for transit of 8.0 hr
12	F11.3	ppm	CDPP rms value for transit of 9.0 hr
13	F11.3	ppm	CDPP rms value for transit of 10.0 hr

(This table is available in its entirety in machine-readable form.)

light curve. Having such a metric is essential to be able to analytically determine the expected MES value of a given transit ($\text{MES} \propto \text{transit depth}/\text{CDPP}$). Here, we investigate the following transit durations: 1, 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8, 9, 10 hr.

To calculate this value, we consider the processed light curve for each target (after EVEREST, both GP fitters, and the harmonic fitter have detrended the light curve). We then inject a transit with depth equal to the standard deviation of the baseline light curve and a period that will reflect $N_{\text{tr}} = 9$ transits within the light curve ($P \approx 8.31$ days). For clarity, this postprocessing injection is independent of the preprocessing signal injections considered in Section 5. The injections discussed in the current section are only intended to measure the residual noise within each light curve. Using a short period with several transits allows us to quickly cover a larger portion of the light-curve parameter space. It also allows us to remain robust to data gaps and systematic within the light curve. If one of the transits were to fall on a data gap, it should only decrease the measure of MES by 6% for that run. Repeated iterations of this simulation will minimize this offset. The transit duration is tuned via the a_{pl} parameter to match the CDPP duration of interest. Then the light curve is examined with TERRA, at a period range corresponding to the injected signal, and the recovered MES value is used to measure the CDPP:

$$\text{CDPP} = (\text{depth}/\text{MES}) \times \sqrt{N_{\text{tr}}} \quad (23)$$

where depth and CDPP are given in units of ppm. Resetting the light curve after each run, these injections are made 50 times with a randomly selected ephemeris at each draw, ensuring sufficient coverage of the time series. To summarize, nine transits are injected into each light curve and this is repeated 50 times, providing 450 injections in each light curve. The recorded CDPP value is then derived using the arithmetic mean of these 50 CDPP measurements. Using nine transits, as done here, allows us to statistically cover the entire phase space of the data for the CDPP duration of 4 hr with only 50 samples, dramatically speeding up our calculations. This process is then repeated for each of the 12 CDPP durations. A machine-readable

version of our table of CDPP values is available. Table 1 shows an example of the contents.

5. Injections

To test the likelihood of a signal being recovered, we inject artificial transit signals into the raw flux data. This is similar to the light-curve-level injections done by Christiansen et al. (2013, 2015) and Burke & Catanzarite (2017b). By doing so, we can then run the flux measurements through the pipeline and count the number of recovered signals. This allows us to measure the detection efficiency of our pipeline.

Our injections utilize the `batman` Python package to create our artificial transits (Kreidberg 2015). When injecting signals, it is important to maximize the number of signals near the MES threshold; this ensures good coverage of the detection efficiency curve (Christiansen et al. 2013). For each available light curve, we inject a transit signal that has a period uniformly drawn from a range of [0.5, 38] days and a value of R_{pl}/R_* drawn from a log-uniform distribution. The log-uniform draw for R_{pl}/R_* is meant to emphasize smaller planets where the signal is near the detection threshold. When injecting signals at the light-curve level, it is important to be cautious about contaminating flux from nearby sources. If the neighboring source contributes a significant amount of flux to the target, the injections can be biased and the pipeline completeness will be artificially improved (planets with smaller radius will appear easier to detect than what reality dictates). To combat this issue, we use the metrics described in Section 3.2 to measure the amount of flux contamination. Instead of sampling from a range [0.01, 0.1] in log space for radius, we sample from $[0.01/\sqrt{\max(F_{\text{Gaia}}^{\text{Tot}}, F_{2\text{MASS}}^{\text{Tot}})}, 0.1/\sqrt{\max(F_{\text{Gaia}}^{\text{Tot}}, F_{2\text{MASS}}^{\text{Tot}})}]$ for each target. This ensures that we do not bias our injection sample. In most cases, the correction for flux contamination is negligible, so the original [0.01, 0.1] sampling space is recovered.

It is important that these signals meet all of the vetting requirements to ensure that we are not injecting planets that will certainly be rejected regardless of the detection capabilities of the pipeline. The most basic cut we make is on signals with fewer than three transits. We uniformly draw the ephemeris (t_0) from a range that will certainly produce three transits within the span of the data. This is uniform with a range of [0, P] for short-period injections (<25 days) and a range of [0, $t_{\text{span}} - 2P$] for longer-period injections. Additionally, it is important that the signals do not trigger an FP flag for an eclipsing binary. In selecting the impact parameter (b) for the artificial planet, we uniformly draw from [0, 1], but such cases where b and R_{pl}/R_* are large will result in an FP flag. We test our randomly drawn values against Equation (11), and both values are redrawn if they exceed this limit. This cycle continues until the b and R_{pl}/R_* values meet this criterion. Drawing impact parameters in this fashion leads to slight bias in our measure of completeness, as we force the injections to meet one of our detection metrics. However, this effect can be easily accounted for as discussed in Section 9.1. For a_{pl} and the limb darkening parameters (u_1 and u_2), we calculate the expected values using the ATLAS model coefficients for the *Kepler* bandpasses tabulated by Claret & Bloemen (2011) and the photometrically derived stellar parameters available on ExoFOP (see footnote 14) (Huber et al. 2016). If such values are not available for the target in question we assume solar values. This assumption is made for less than 1% of the targets

and will only impact the inferred limb darkening parameters, avoiding the introduction of serious biases to our targets. The updated stellar parameters discussed in Section 8.1 were not available when this processing occurred. In testing, we found this use of older stellar parameters had no effect on our ability to recover signals. All final inferred planetary parameters were calculated using the most up-to-date stellar values (Hardegree-Ullman et al. 2020).

The injected raw flux data are then processed through our pipeline (EVEREST, TERRA, and EDI-Vetter). We consider the recovery of a signal as a TCE and as a vetted candidate. Looking at this recovery fraction in terms of period and radius (Figure 6), we see that the vetting metrics in EDI-Vetter do not induce a significant loss in the recovery fraction ($\sim 10\%$; which is comparable to that seen by the `Robovetter` (Thompson et al. 2018) for *Kepler* as seen in Figure 7). A large portion of the vetting loss comes from edge cases ($P \sim 0.5$ days, $P \sim 38$ days, and $R_{\text{pl}}/R_* \sim 0.10$), where statistical fluctuations can easily push the parameters beyond the threshold of our vetting limits. In general, we find that the pipeline has the most difficulty recovering transits with long periods and small R_{pl}/R_* , and the pipeline is most adept at recovering signals with short periods and large R_{pl}/R_* . This expected trend indicates that the detection efficiency of our pipeline is a function of MES.

It is essential that we understand the noise (CDPP) within each light curve so that we can analytically determine the expected MES value for each injection. This will allow us to accurately calculate the recovery fraction throughout the MES range in question. We use the following equation to determine this parameter:

$$\text{MES} = \text{MES}_{\text{Cor}} \times \frac{\text{depth}}{\text{CDPP}_{t_{\text{dur}}}} \times \sqrt{N_{\text{tr}}} \quad (24)$$

where $\text{CDPP}_{t_{\text{dur}}}$ is the CDPP value corresponding to the injected transit duration and MES_{Cor} is the correction factor that matched the analytic function to the MES being detected by the pipeline (MES_{Det}). Since the CDPP values derived in Section 4 are discrete, $\text{CDPP}_{t_{\text{dur}}}$ is an interpolation between the nearest values of duration. Cases where $t_{\text{dur}} < 1$ hr or $t_{\text{dur}} > 10$ hr are not extrapolated, but rather assigned the nearest CDPP value. To calculate MES_{Cor} we look at the most well-behaved injections ($\text{MES}_{\text{Det}} < 200$ and $b < 0.1$). These restrictions limit the amount of MES smearing (Burke & Catanzarite 2017b) that our injections will experience. This type of smearing occurs because the detected MES value is drawn from a distribution of values centered around the theoretical MES value. Signals with a larger MES or high impact parameter are prone to wider deviations from the theoretical MES value, making it more difficult to estimate the theoretical values. To minimize the potential bias that this smearing effect can impose, we take the median value of $\text{MES}_{\text{Det}}/\text{MES}$, which we find to be 0.880, as our MES_{Cor} value. This indicates that our analytic function will overestimate the MES value by 12% without this correction factor. This factor is much more significant than the one found for *Kepler* (4.4%; Christiansen 2017), but is not surprising because the *K2* data are more unruly. It is important to remember that even with this correction in place, the analytic MES value may differ from MES_{Det} on any given signal;

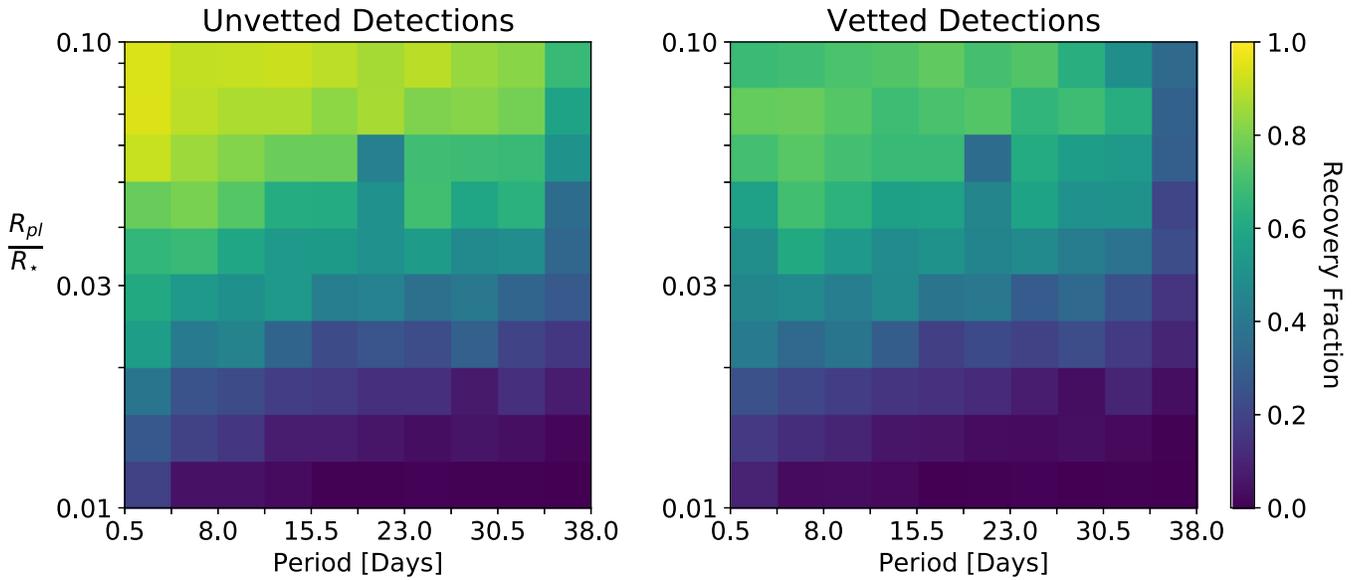


Figure 6. The recovery fraction of injected planets as a function of period and radius. Each bin is sampled uniformly (~ 250 injections each). The left plot shows the recovery fraction for injected planets detected as a TCE. The right plot shows the recovery fraction for injections that were detected as a TCE and also labeled as a candidate by EDI-Vetter. The anomaly seen at $P = 20$ days and $R_{pl}/R_* = 0.05$ appears to be statistical fluctuation where an abnormally large fraction of targets with high CDDP were selected for injection. We have thoroughly checked our pipeline and find no reason why a signal in this specific range would have a unique recovery fraction.

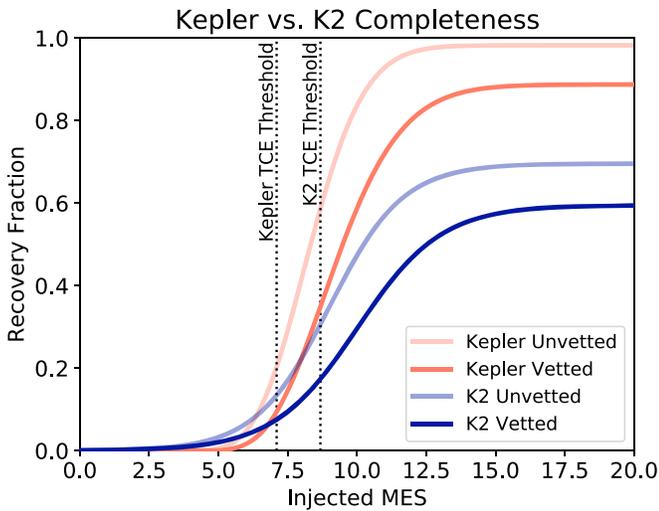


Figure 7. A comparison of the completeness functions for *Kepler* and *K2*. The light peach and peach lines represent the Gamma functions for the *Kepler* DR25 TCE (Christiansen 2017) and the Robovetter (Coughlin 2017) completeness respectively. The light blue and blue lines represent the logistic functions for the *K2* TCE and EDI-Vetter completeness functions derived in this study. All functions have been fit using injections across all ranges for which periods are available.

the analytic MES value represents an average over the entirety of noise within the light curve, while MES_{Det} is a discrete measurement.

As done in previous completeness measurements, we consider the separation of M dwarfs from FGK stars. Typically M dwarfs have more correlated stellar noise and produce, on average, a decreased completeness function. We begin this separation with our initial list, which contains 25,040 stellar targets. Removing targets with apertures that are too large or have incomplete data sets, we retain 24,949 targets. We then use the stellar parameters provided by Hardegree-Ullman et al.

(2020) to establish the spectral type classifications and $\log g$ values for our targets. To remove giants from our sample we eliminate targets with $\log g < 4$ (17,045 targets remain). We also remove targets that exhibit noisy light curves ($CDPP_{8hr} > 1200$ ppm). We select this limit of CDDP because it roughly corresponds to 5σ from the median CDDP value of all targets. This cut leaves us with 15,698 targets. We now separate the sample into FGK dwarfs using the spectral classification provided in Hardegree-Ullman et al. (13,046 targets) and M dwarfs (2652 targets). We initially consider these two populations separately in our completeness calculation. However, we find no statistically significant difference in the detection efficiency between the two spectral groups. This is possibly caused by the small sample of M dwarfs available in Campaign 5, and we will continue to test this potential separation of stellar spectral type.

Looking at bins of width 1 MES, we calculate the fraction of recovered injections as a function of MES. We use the analytic MES values found using Equation (24) for both recovered and lost signals to ensure consistency. We also look to see what fraction is recovered once the signals have been processed through EDI-Vetter (Figure 8). To account for inner bin fluctuations, and provide a more precise measure of recovery, we use the kernel density estimator (KDE) for fitting. To calculate the KDE, we use a uniform kernel with a width of 1 MES. This method essentially recalculates the recovery rate at each shift of 0.05 MES. The uncertainties are calculated assuming binomial statistics, because the options are binary (the signal is either recovered or not recovered). The measured KDE values are then used to fit a logistic function ($f(x)$) to the data:

$$f(x) = \frac{a}{1 + e^{-k(x-l)}} \quad (25)$$

where a , k , and l are all tunable parameters. In testing, we tried several functions, including the commonly used Gamma function (Christiansen et al. 2015; Christiansen 2017), but found that the logistic function provides the best fit with the detection efficiency

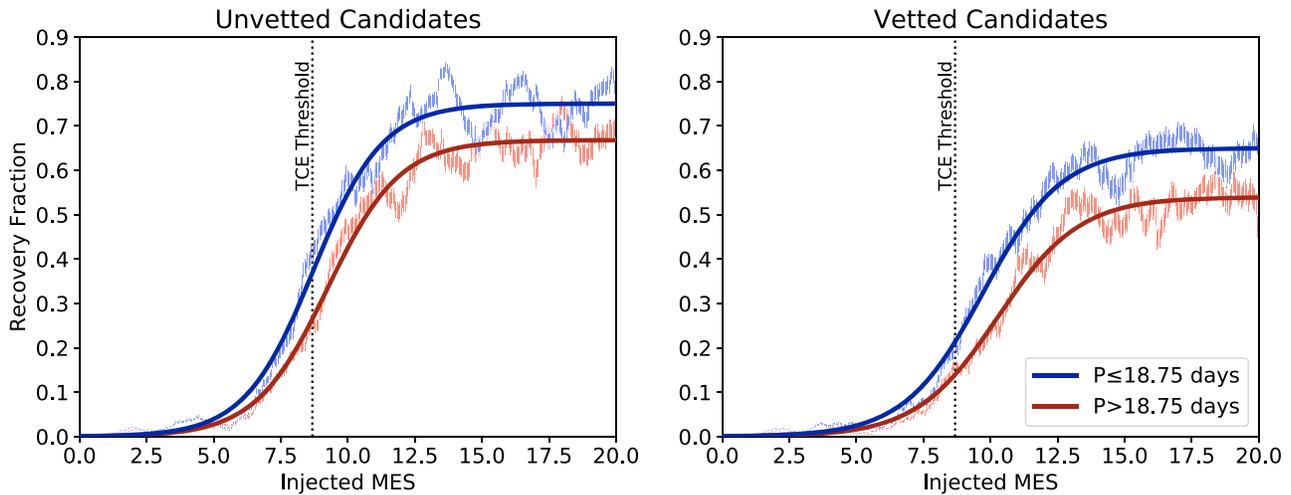


Figure 8. Injected signal MES as a function of the recovery fraction. The left plot shows the recovery function for planets detected as a TCE. The right plot shows the recovery function for planets detected as a TCE and also deemed a candidate by EDI-Vetter. The blue and red vertical lines represent the kernel density estimator (KDE) using a uniform kernel with a width of 1 MES. The height of these lines represents the binomial uncertainty of each measurement. The dark blue and red lines illustrate the best fit logistic function given by the parameters in Table 2. We choose to separate the data at large periods to account for the decreased signal recovery from injections with only a few transits. All injections with periods < 18.75 days will have at most three transits.

of our pipeline. To fit the data we maximize the binomial log-likelihood function ($\log L(x|\text{found}, \text{lost})$):

$$\log L(x|\text{found}, \text{lost}) \propto (\text{found}) \times \log(f(x)) + (\text{lost}) \times \log(1 - f(x)) \quad (26)$$

where “found” represents the number of recovered injections in a given KDE measurement and “lost” represents the number of lost signals in a given KDE measurement. The results of our fitting procedure are provided in Table 2.

It is apparent that the *K2* pipeline is far less efficient than the previous *Kepler* study. Comparing the unvetted candidates (70% efficient at high MES) with the results of Christiansen (2017) (94% efficient at high MES), we can see a 24% drop in the recovery rate from *Kepler* to *K2* in Figure 7. This is significant, but expected because *K2* data have far more systematic noise issues, causing transit detection to be more difficult. However, understanding this deficiency allows us to account for the loss when making measurements of occurrence.

We note that a more thorough test would be to inject the transits at the pixel level as done by Christiansen (2017) for the *Kepler* pipeline, but this process is far more computationally expensive and only provided a small correction (an increase of 2%) to the *Kepler* detection efficiency. We will consider such undertakings in future studies of the *K2* completeness.

6. Reliability

To assess the reliability of the catalog against instrumental FPs, we flip the processed light curves upside-down and run them through TERRA and EDI-Vetter. Since the light-curve inversion ensures that no true astrophysical transit exists within our data set, counting up the number of false signals that make it through the vetting process gives us a handle on the reliability of our candidates. Our pipeline finds 6145 TCEs in the inverted light curves (compared to 5452 found from the original processing), and only four (0.07%) of these make it through the EDI-Vetter testing.

To determine how reliable our pipeline is, we use the metrics laid out in Equations (3)–(8) of Thompson et al. (2018). First, we must calculate how efficiently FPs are being classified by

Table 2
The Logistic Function Parameters that Best Model the MES Recovery Fraction

	a	k	l
Unvetted			
$P \leq 18.75$ days	0.75023	0.7769	8.7199
$P > 18.75$ days	0.6680	0.7434	9.2445
Vetted			
$P \leq 18.75$ days	0.6499	0.6813	9.7288
$P > 18.75$ days	0.5394	0.6572	10.2787

Note. These values were obtained by maximizing the binomial likelihood function for the KDE measurements provided in Figure 8.

the pipeline (E), using the inverted light-curve results:

$$E \approx \frac{N_{\text{FP}_{\text{inv}}}}{T_{\text{FP}_{\text{inv}}}} \quad (27)$$

where $T_{\text{FP}_{\text{inv}}}$ is the total number of TCEs found in the inverted light curves and $N_{\text{FP}_{\text{inv}}}$ is the number of TCEs found that were vetted and accurately assigned FP status. Since our test of inverted light curves only produced four falsely identified candidates, small-number statistics need to be considered. To account for these small-number statistics, we add one additional false alarm signal to each measure of E so that it becomes

$$E \approx \frac{N_{\text{FP}_{\text{inv}}}}{T_{\text{FP}_{\text{inv}}} + 1}. \quad (28)$$

This addition is motivated by the expectation value for λ given N measurements from a Poisson distribution using Bayes’ theorem:

$$\begin{aligned} P(\lambda|N) &\propto P(N|\lambda) * P(\lambda) \propto \lambda^N e^{-\lambda} * 1 \\ &\propto \text{Gamma}(\text{shape} = N + 1, \text{scale} = 1) \\ \langle P(\lambda|N) \rangle &= \langle \text{Gamma}(N + 1, 1) \rangle = N + 1 \end{aligned} \quad (29)$$

where Gamma represents a Gamma distribution. However, this means we are likely underestimating the efficiency

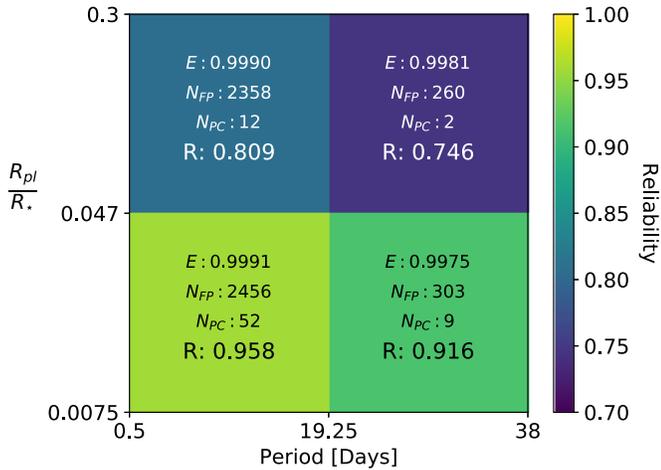


Figure 9. A display of the reliability of our pipeline. The N_{FP} values correspond to the number of FPs identified in each bin. The N_{PC} values indicate the number of planet candidates found in each bin. The R value is the expected reliability given the number of candidates and FPs using Equation (30). The bins have been divided at the midpoint of the ranges considered in this search. The R_{pl}/R_s cut is at the midpoint in log space.

of our FP identification. Thus, our reliability values should be considered a lower limit that will likely increase as we increase our FP count with data from other *K2* Campaigns. Overall, our pipeline finds $N_{FP_{inv}} = 6141$ and $T_{FP_{inv}} = 6145$; using Equation (28) our overall pipeline E is 0.9992. In Figure 9 we report the E values calculated for each region of parameter space.

Using the number of planet candidates (N_{PC}) and the number TCEs given FP status (N_{FP}) in the non-inverted light curves, we can determine the reliability fraction (R) of our data set:

$$R = 1 - \frac{N_{FP}}{N_{PC}} \left(\frac{1 - E}{E} \right). \quad (30)$$

We provide a coarse binning of these reliability metrics in Figure 9. Overall our pipeline finds a reliability fraction of 0.942, indicating that we should only expect ≈ 4 FPs to be contaminating our candidate catalog. Again, small-number statistics are largely at play here, and it is very possible that our true reliability is even greater than the values we have reported. This is slightly less reliable than the *Kepler* pipeline, which finds an overall reliability of 0.97 for its all-stars measure (Thompson et al. 2018). However, a more appropriate comparison would consider the parameter space where a similar number of transits occur (periods > 10 days for *Kepler*). Using Figure 8 of Thompson et al. (2018) and knowing that 18,660 TCEs were detected with periods beyond 10 days, we estimate this reliability to be roughly 0.95 for the *Kepler* data set in a comparable parameter space. While this value is still higher than the measurement presented by our pipeline, we remind the reader that a conservative estimate of reliability has been provided, which will likely improve as we increase our sample.

Testing reliability by inverting the light curves assumes that the FPs are symmetric and will persist upon inversion. One method of testing this assumption is by examining the period distribution of the TCEs. In Figure 10 we compare the TCE distribution to the inverted TCE distribution. Overall, the

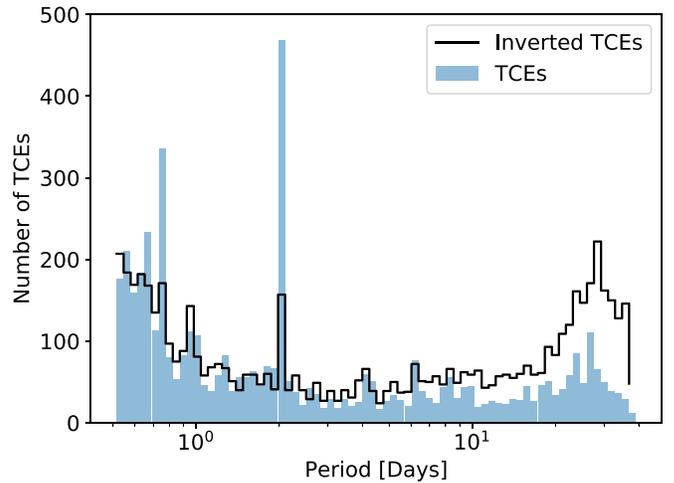


Figure 10. A plot of the TCE distribution and the inverted light-curve TCE distribution.

distributions provide a reasonable match. We find some underrepresentation in the inverted TCEs near the Resat period (2 days) and the third thruster harmonic (0.75 days), indicating that these systematics are slightly asymmetric. Additionally, the long-period TCEs are slightly overrepresented in the inverted light curves. We acknowledge that these mismatches could modify our measure of reliability. Additional light curves from other campaigns will help to improve this measure, and other methods such as data scrambling (see Section 2.3.2 of Thompson et al. 2018) may be considered in future studies.

7. Window Function

The window function gives the probability (prob) that a certain period (P) will meet the requirement of a minimum of three transits for TCE consideration within the span (t_{span}) of the available light curve (von Braun et al. 2009; Ballard et al. 2010). This metric is an essential ingredient for measurements of occurrence (Burke et al. 2015; Bryson et al. 2019; Zink et al. 2019a), where all aspects of detection probability must be considered. If the data are seamless, without any masking, the equation can be found analytically:

$$\begin{aligned} \text{prob} &= 1; & P < t_{span}/3 \\ \text{prob} &= \frac{t_{span}}{P} - 2; & \frac{t_{span}}{3} \leq P \leq \frac{t_{span}}{2} \\ \text{prob} &= 0; & P > t_{span}/2 \end{aligned} \quad (31)$$

where t_{span} for Campaign 5 is nominally 74.82 days. However, this problem becomes more difficult to solve when considering masked cadences within each light curve. While some cadences are masked globally (see Section 2.4.1), often the masking is done on an individual per-target basis, in an attempt to remove outliers. To ensure accuracy, we directly measure this probability for each target light curve, similarly to Burke & Catanzarite (2017a) for *Kepler* DR25.

Looking at each light curve, we used the corresponding masked cadences (those that were masked prior to the TCE search process) to create the gaps in our light curve. Starting at 18 days, we take steps of 0.1 days and test the probability of each period yielding three transits. It is important here to consider how the transit duration will affect the number of available cadences in each transit. Longer transits will have a

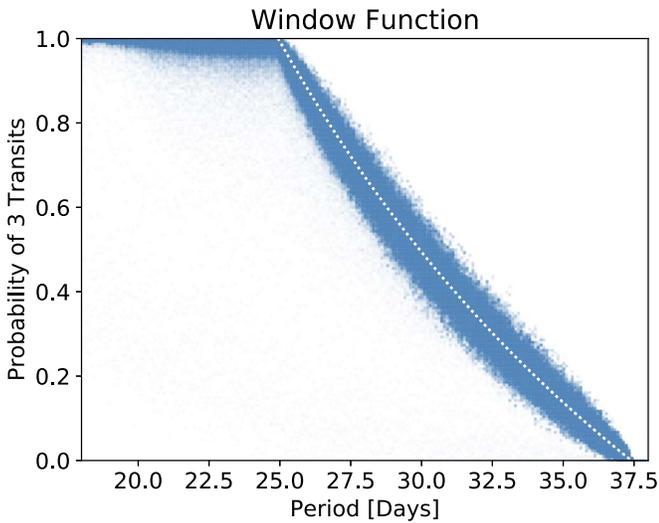


Figure 11. The measured window function for all targets in Campaign 5. The white dotted line represents the analytic function (Equation (31)) expected for this campaign, assuming $t_{\text{span}} = 74.84$ days. We see that a large fraction of the targets appear to exceed the functional probability, which represents the best-case detection scenario. This is due to the increased t_{span} for a limited number of targets, but more importantly, to statistical fluctuations in our Monte Carlo simulation. The increased dispersion seen near $\text{prob} = 0.5$ is the signature of a binomial distribution, caused by statistical fluctuations. The data behind this figure are available in machine-readable format.

(The data used to create this figure are available.)

higher likelihood of avoiding serious masking. To marginalize over this parameter, we sample from our empirical sample of planet candidates with replacement. This sampling is biased toward planets with longer transits (which are easier to detect). To account for this, we do not sample uniformly, but rather with a $1/\sqrt{t_{\text{dur}}}$ probability (as MES scales with $\sqrt{t_{\text{dur}}}$). This gives more weight to shorter transits. This weighted sampling produces a median t_{dur} of 2.20 hr with a minimum of 1.05 hr and a maximum of 8.38 hr. A more thorough metric would be to create a 2D grid, stepping through both period and t_{dur} values, but we found little variation when changing the transit duration, and have settled on a simple measure using only the planet period. At each step of period, we sample 160 times (roughly 2% accuracy for a binomial distribution), selecting a new t_{dur} and ephemeris (t_0) at each iteration. We then count up the number of times three or more transits occurred within the span of the data set. We only consider a transit as observed if at least 50% of the transit was unmasked. We then record the probability of three transits occurring at this period and move on to the next step in period space. Figure 11 shows all of the measured window functions. It is apparent that the masking is spread uniformly though the light curves, minimizing the deviation from the analytic Equation (31). If significant gaps exist in the light curves, the measurements produce large deviations from the functional form. To a first-order approximation, the analytic equation provides a good measure of the window function. For a more detailed calculation, we provide the measured window function for each target as the data behind Figure 11.

8. Candidates

Using our described pipeline, we detected 75 planet candidates in the 15,698 C5 light curves searched. The parameter estimations

made by the pipeline are performed with a focus on computational speed. With our final list of 75 candidates, we can afford to let our MCMC sampler more thoroughly explore the likelihood surface. We use `emcee` (Foreman-Mackey et al. 2013) with 100 semi-independent walkers, a burn-in of 500 steps, and a parameter space coverage of 500 steps (50,000 test samples). In this fit we allow period (P), impact parameter (b), ephemeris (t_0 ; with a zero offset of 2,454,833 days), radius ratio (R_{pl}/R_{\star}), and semimajor axis (a) to vary at each iteration. We assume circular orbits for all of our planets and use the limb darkening parameters derived from the existing stellar parameters (we discuss these non-variable transit parameters in Section 5). Table 3 shows an example of our results. For each candidate, we also make a snapshot of each transit, which includes the full light curve, the folded transit, and the residuals when the transit model is removed. We provide an example of this snapshot for planet candidate EPIC 211816003.01 in Figure 12. We provide our full list of candidate snapshots on ExoFOP (see footnote 14).

As mentioned in Section 5, possible flux contamination can cause an underestimation of the transit depth. We use the parameters described in Equation (11) to correct for detected contamination. All R_{pl}/R_{\star} parameters are multiplied by the flux contamination factor $\sqrt{\max(F_{\text{Gaia}}^{\text{Tot}}, F_{\text{2MASS}}^{\text{Tot}})}$, increasing the ratio accordingly. We note that this factor only accounts for flux contamination detected by either *Gaia* or 2MASS. It is likely that other transit dilution exists from bound binary companions (Ciardi et al. 2017; Furlan et al. 2017; Matson et al. 2018). High-resolution follow-up is essential to rule out such cases for our candidate list. Additionally, signal processing has the potential to artificially decrease the transit depth. We discuss this issue in greater detail in Section 8.2. We note here that our listed candidates do not reflect any type of systematic correction; we leave such a procedure to be carried out at the reader’s discretion.

Within our candidate sample we find seven systems with more than one detected planet. System multiplicity is often a good indicator that the planets are real and not FPs (Lissauer et al. 2014), thus we focus our discussion on these candidates. One target, EPIC 211428897, contains four planet candidates. Three of the candidates were previously noted (Dressing et al. 2017; Petigura et al. 2018). Alongside Kruse et al. (2019), we find an additional candidate at a period of 6.265 days. We also verify the claim of Kruse et al. (2019) for two new multi-planet systems: EPIC 211413752 and EPIC 212072539 both had a single known candidate (Petigura et al. 2018), but we were able to extract additional planets in both of these systems, increasing the number of known multi-planet systems. In Section 8.1 we briefly discuss the how our multi-planet candidates are distributed among stellar spectral type.

We suspect that several of the single-planet systems likely have additional planets that either do not transit or were not detected because of the issues regarding detection order pointed out by Zink et al. (2019a) for *Kepler*. Many of these same detection features exist in this pipeline, as discussed in Section 2.3. An example of this is EPIC 212164470 (Barros et al. 2016; Pope et al. 2016; Mayo et al. 2018; Petigura et al. 2018), where we only recovered the transit with a period of 7.81 days and did not recover the smaller MES signal at 1.74 days (Mayo et al. 2018). In later iterations of this pipeline, we will test the detection efficiency of multiplicity by injection of several planets into each light curve, allowing us to directly

Table 3
An Example of the Planet Candidate Table for Campaign 5 Available for Download in Machine-readable Format

Candidate	P (days)	R/R_*	R (R_{\oplus})	t_0 (BJD)	b	t_{dur} (hr)	...
211562654.04	2.15290 ± 0.00015	0.01151 ± 0.00062	1.23 ± 0.11	2307.4643 ± 0.0035	0.47 ± 0.14	2.03 ± 0.21	...
211711685.01	15.4624 ± 0.0023	0.01273 ± 0.00050	1.28 ± 0.06	2310.8848 ± 0.0054	0.47 ± 0.17	3.94 ± 0.55	...
211918985.01	9.4892 ± 0.0040	0.0390 ± 0.0018	4.40 ± 4.35	2316.0782 ± 0.0078	0.26 ± 0.19	8.38 ± 1.06	...
211942755.01	19.7288 ± 0.0034	0.0285 ± 0.0013	2.22 ± 0.13	2322.8335 ± 0.0038	0.688 ± 0.056	2.70 ± 0.21	...
211953244.01	29.7236 ± 0.0051	0.01547 ± 0.00097	1.44 ± 0.10	2307.3432 ± 0.0056	0.67 ± 0.18	4.07 ± 0.90	...
211958340.01	1.465460 ± 0.00028	0.2693 ± 0.0074	42.49 ± 1.72	2307.17786 ± 0.00082	0.468 ± 0.058	3.23 ± 0.19	...
212020330.01	1.65348 ± 0.00039	0.01430 ± 0.00059	1.87 ± 0.14	2307.8556 ± 0.0051	0.78 ± 0.12	2.69 ± 1.20	...
212119244.01	1.07692 ± 0.00023	0.00772 ± 0.00045	1.64 ± 1.62	2307.1496 ± 0.0055	0.41 ± 0.25	2.02 ± 0.47	...

Note. Listed are the new planet candidates detected in this paper.

(This table is available in its entirety in machine-readable form.)

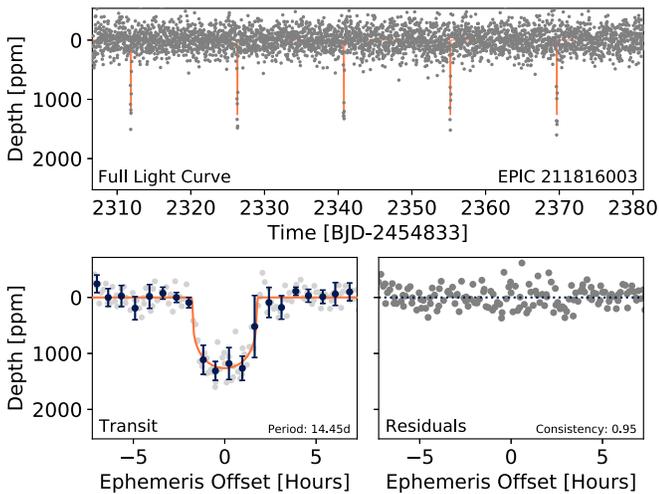


Figure 12. An example of a fully vetted planet candidate. The gray points represent the processed flux measurements and the navy points show the binned average. Here we choose a bin width that ensures five bins exist within the transit duration. The orange line represents the best fit transit model.

measure the role that detection order plays in transit recovery. However, such a task is beyond the scope of this paper.

One reason we lose multiple planet signals is the limited range of our period search. We limit our pipeline to a range of 0.5–38 days. Previous surveys have found two systems that have multiple transiting planets with periods less than 0.5 days (EPIC 211305568, Dressing et al. 2017; EPIC 211562654, Mayo et al. 2018). Upon visual inspection of the light curve, it appears that these candidates were found with our TCE search but then rejected by our vetting requirement of $P \geq 0.5$ days. Weakening this vetting requirement may yield more planets, but we found in practice that it will dilute the purity of our sample (Section 2.3). We note that a large fraction of the targets in Campaign 5 are also available in either Campaign 16 or 18. By stitching the data from multiple campaigns together, the available data span can be extended to potentially find periods beyond our 38 day limit. An example of this is EPIC 211611158, which has a known planet at a period of 52.71 days (Mayo et al. 2018). It should be noted that this planet was not found by stitching, but rather by relaxing the requirement of three transits. Currently, there has been no effort to combine overlapping campaigns in the manner discussed. We hope to develop the ability to stitch together campaigns in future iterations of this pipeline.

One system of known planets was completely lost by our pipeline, EPIC 212157262. This system was detected and also confirmed by Mayo et al. (2018), who used the K2SFF (Vanderburg & Johnson 2014) processed light curves. Some of the planets in this system produce signals near the detection threshold and were not detected in our pipeline TCE search. Using different detrending algorithms can yield slightly different results, which may help push some signals above this threshold while moving others below it. Using all the different detrending algorithms simultaneously can help minimize such loss (Kostov et al. 2019), but requires a serious computational cost to create the completeness and reliability metrics available here. Thus, we have only used the EVEREST detrending algorithm, and we accept the potential planet loss. Additionally, two of the planets (K2-187 d, e; Mayo et al. 2018) have strong signals and should not be near the detection threshold. In fact, these two planets were detected in our injection simulation. This loss is caused by the indeterministic nature of the EVEREST software, meaning that the same flux data can produce two different processed light curves on different iterations. One way to combat this problem would be to process all of the light curves twice and take the case with the lowest CDP, but such a task would significantly increase the amount of computational time required and is not done here. We run each light curve through EVEREST twice; in doing so we acknowledge that individual injected light curves may produce better or worse CDP than our nominal observational light curves. However, when considering the overall injection population this effect should cancel out. Such issues are folded into our completeness measure in Section 5. Future iterations of our pipeline will attempt to minimize this effect by forcing EVEREST to abide by a fixed seed.

Our search pipeline is able to identify nine previously unknown candidates (see Table 3). These new candidates are mostly planets with a small radius ($R < 3 R_{\oplus}$) that were not detected by previous searches, which used higher MES thresholds. One of these candidates extends the multiplicity of known systems (EPIC 211562654). Almost all of the new candidates have *Kepler* band magnitudes < 13 , with the exceptions being EPIC 211918985.01 and EPIC 211942755.01, which are 16.4 and 14.1 mag respectively. The brightness of these new candidates makes them excellent targets for follow-up validation.

One candidate (EPIC 211958340.01) appears to be suspiciously large for a planet, highlighting the limitations of our stellar agnostic vetting metric (Equation (11)). This candidate’s host star has a relatively large radius ($1.44 R_{\odot}$), where

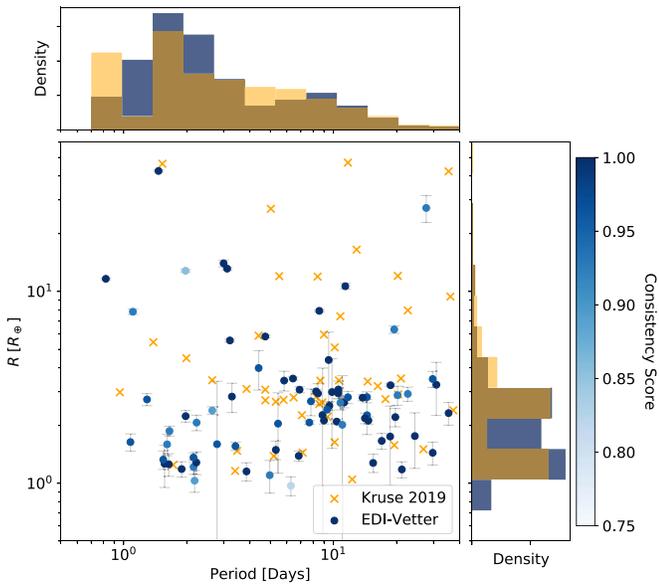


Figure 13. A plot of the candidates resolved by our pipeline. The dots (EDI-Vetter) correspond to the candidates detected, with a color scaling to show the consistency score of each candidate (Section 3.13). The \times symbols represent the candidates not found by our pipeline but noted by Kruse et al. (2019). The histograms show the density of the found planets for both period and radius by our pipeline in blue and the entire Campaign 5 catalog of Kruse et al. (2019) in orange.

eclipsing binaries have the potential to produce small dips in the light curve. We suggest that a study of occurrence imposes an upper limit on radius to avoid these potential astrophysical FPs. Similar contaminants can be found in the *Kepler* pipeline candidate list.

Overall, our pipeline detected 75 planet candidates. Comparing these to known candidates from Kruse et al. (2019), who manually vetted transit signals, we recover 51% of the known planet candidates within our search limits on period and radius ratio. This manually vetted catalog contains the largest yield of *K2* candidates to date and also used the EVEREST software to remove spacecraft systematics, thus we use Kruse et al. (2019) for comparison in Figure 13. It is important to remember that while Kruse et al. (2019) may provide a more complete list of the candidates from Campaign 5, our catalog of candidates has been uniformly vetted, eliminating potential confirmation biases. Comparing to the systems of multiple planets, where FPs are less likely, we recover 60% of the known multi-planet system candidates. An alternative comparison can be made for Petigura et al. (2018), who also used TERRA to search for transit signals in Campaign 5 of *K2*. We recover 70% of the planets in the manually vetted catalog of Petigura et al. (2018) that fall within our search limits on period and radius ratio. Two other large transit searches have been performed on Campaign 5, by Dressing et al. (2017) and Pope et al. (2016). We find that our candidates list overlaps with 51% of each of these previous studies within our search limits on period and radius ratio.

There are several reasons why we fail to recover 100% of the known planet candidates. As mentioned previously, using different detrending algorithms will yield different results and cause us to lose some planets that a different algorithm would otherwise allow us to detect; 37% of the lost planets were never detected as a TCE by our pipeline. Additionally, our vetting

metrics are harsher than some studies, which take an “innocent until proven guilty” philosophy toward planet candidacy. We aim for reliability, so we are willing to remove real signals at the benefit of ensuring fewer FPs. The eclipsing binary metric (Equation (11)) of our vetting removes many of the potential candidates found by other studies, accounting for 17% of the lost planets. The difference in even and odd transit depths may cause a rejection when one of the transits is poorly detrended (Section 3.5), leading 13% of the rejections of known planet candidates. The outlier metric (Section 3.3), meant to remove signals dominated by the chance alignment of several outliers, can misclassify noisy transit signals, removing 13% of the lost planets. The remaining 20% are lost due to single triggers of various other metrics in EDI-Vetter.

8.1. Stellar Parameters

The measurements of planet radius available in Table 3 were calculated using the values of stellar radius given by Hardegree-Ullman et al. (2020). These parameters were inferred using data from photometric *g*, *r*, *J*, *H*, *K*, *Kepler*, and *Gaia* bands in combination with the Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST; Su & Cui 2004) spectra. Using the well defined parameters (spectral type, T_{eff} , $\log g$, [Fe/H]) available from the LAMOST data, a random forest algorithm (Pedregosa et al. 2011) was used to assign these parameters to each of the *K2* targets based on the photometry for targets without spectroscopic measurements. Bolometric luminosities were computed using *K*-band magnitudes and *Gaia* parallaxes, from which stellar radii were calculated using the Stefan–Boltzmann law. In Table 3 we provide the transit parameters in such a way that stellar and planet parameters can be updated as new information becomes available. Using this updated stellar classification, our list of candidates includes 58 planets around FGK stars and 17 that are hosted by M-dwarf stars. We find eight planets in multi-planet systems around M dwarfs and nine in multi-planet systems around FGK dwarfs. This indicates that nearly 50% of the M-dwarf candidates are part of a system of detectable planets, which is consistent with the occurrence rate of compact multiple planets of 44% from Hardegree-Ullman et al. (2019). This could be due to the fact that smaller planets are more easily detected around M dwarfs, increasing the multiplicity of detectable planets. With such a small sample (75 candidates) it remains difficult to make strong claims about the population parameters of the planets of M dwarfs versus FGK dwarfs, but overall they both follow similar distributions of period and radius. A more thorough investigation of population parameters will be left for a future study.

8.2. Planet Radius

The *K2* data set must undergo significant processing (Section 2) before the transit may be fit. This extensive manipulation may lead one to question the accuracy of the extracted parameters. The feature that experiences the largest risk of modification is the transit depth. The EVEREST software, the GP detrenders, and the harmonic fitter all have the potential to remove some of the transit depth. Fortunately, injecting signals at the light-curve level (as done in Section 5) affords us the opportunity to measure the transit depth before and after processing.

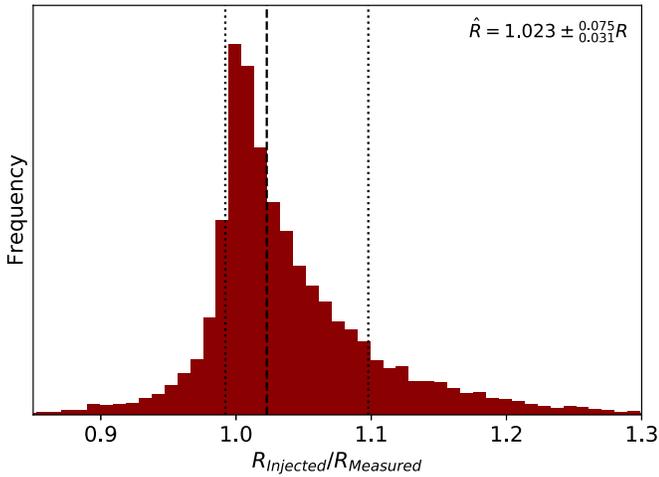


Figure 14. A histogram of the ratio of injected to recovered planet radius with bin widths of 0.01. The detrending and processing of the light curves on average leads to a 2.3% reduction in the inferred planet radius. The dashed line represents the median ratio value and the dotted lines mark the 68% quartiles. This shows that the offset has a statistical significance of $<1\sigma$.

We focus on the planet radius, because this parameter has the largest impact on the transit depth. In Figure 14 we present the ratio of expected radius and recovered radius after processing. We find an expected offset of 2.3% for the planet radius. This means that, on average, our pipeline will underestimate the planet radius by 2.3%. In comparison, the *Kepler* pipeline considered the offset in measured MES and detected MES (0.4%); transforming this metric to radius, the expected offset for *Kepler* is -0.19% for the recovered planets (Christiansen et al. 2013). Unsurprisingly, the *K2* pipeline is far more prone to reduction in transit depth. However, both of these offsets (*Kepler* and *K2*) are smaller than the average uncertainty in planet radius ($\sim 10\%$; Berger et al. 2018). Knowing that this offset falls within the normal uncertainty in radius, Figure 14 shows that the mode of this offset distribution is located at 1; we do not apply this correction to our candidate parameters. We leave such a procedure to be carried out at the reader’s discretion. However, we do incorporate this potential offset into our measure of radius uncertainty. To account for the uncertainty expected by the radius offset (σ_{Off}), we use the following operation:

$$\sigma_{\text{Off}} = 0.023 \times R \quad (32)$$

where R is our best estimation of the planet radius.

We also investigated the possible difference in this ratio distribution when separating long-period injections (>10 days) from short-period injections (<10 days). We expect that planets with more available transits might be less prone to this type of depth reduction, but we find no significant difference between these two distributions. Therefore, we only focus on the combined population.

Beyond the radius offset mentioned above, we also want to consider our overestimation of flux contamination (Section 3.2). We assumed a 1D model for flux contamination, but a 2D model would produce less contamination ($\sim 1.88\times$ in the worst-case scenario). To address this issue, we fold the potential

overestimation (σ_F) into our measure of radius uncertainty:

$$R = R_{\text{fit}} \times R_* \times \sqrt{F^{\text{Tot}}}$$

$$\left| \frac{\partial R}{\partial F^{\text{Tot}}} \right| = \frac{R}{2 \times F^{\text{Tot}}}$$

$$dF^{\text{Tot}} = \frac{F^{\text{Tot}} - 1}{1.88}$$

$$\sigma_F = \left| \frac{\partial R}{\partial F^{\text{Tot}}} \right| dF^{\text{Tot}} = R \frac{F^{\text{Tot}} - 1}{3.76 \times F^{\text{Tot}}} \quad (33)$$

where R_{fit} is the radius ratio found using our MCMC sampler, R_* is the radius of the stellar host, and F^{Tot} is the ratio of the total flux to the flux of the target star. If F^{Tot} is one, σ_F is zero because no contamination was found, and therefore no correction is warranted. The uncertainty contribution from this parameter is very small ($\sigma_F/R \sim 6.9 \times 10^{-6}$). Our combined estimation of the uncertainty in planet radius is as follows:

$$\sigma_R = \sqrt{\sigma_{\text{fit}}^2 + \sigma_*^2 + \sigma_F^2 + \sigma_{\text{Off}}^2} \quad (34)$$

where σ_{fit} is the uncertainty measured by our MCMC sampler for R_{fit} ($\sigma_{\text{fit}}/R \sim 0.040$) and σ_* is the uncertainty measured for the stellar radius ($\sigma_*/R \sim 0.06$). In almost all cases σ_{fit} and σ_* dominate the uncertainty, but the other parameters provide a slight increase to account for the issues they represent. Overall, we find a median (σ_R/R) of 0.16 for our list of planet candidates.

The radius distribution recovered by our pipeline shows some hints of the gap in radius near $2 R_{\oplus}$ (Fulton et al. 2017), shown by the double peak in the blue radius–density plot of Figure 13. The improved stellar parameters available in this study allow this gap to be visible even with our small sample size of 75 candidates. Evidence for this gap in the *K2* data has also been acknowledged in Mayo et al. (2018) and Kruse et al. (2019). However, a more detailed accounting of completeness is necessary to verify such claims. A first-order completeness correction, using post-processing injections, was carried out by Cloutier & Menou (2019), combining the *K2* and *Kepler* data to verify the gap in low mass star systems. However, we leave a complete analysis using light-curve level injections for future studies.

9. Conclusions

We have provided the results from our pilot study, in which we created a uniform planet candidate catalog for *K2* Campaign 5. The results of this study are intended to help generate a population of planets that can be used to measure the underlying exoplanet population for this campaign out to 38 days. We also show that our methodology can be applied to the remaining *K2* campaigns C11–C18, where the number of known candidates is much lower and cannot provide sufficient data for tuning of the pipeline.

Our candidate sample includes 75 planets with seven multi-planet systems (five double, one triple, and one quadruple planet systems), and nine new candidates. We have successfully recovered 51% of the known planet candidates (Kruse et al. 2019) within the range of our survey, and can uniformly measure the completeness and reliability of our sample.

We also introduce our vetting software (EDI-Vetter), which has been optimized to address issues specific to the *K2*

data set. This software builds upon the metrics of *Robovetter* (Thompson et al. 2018), which was used on *Kepler* DR25. We provide several new features that allow us to produce a reliable candidate catalog for *K2*. The philosophy of *EDI-Vetter* is to minimize the number of FPs, even if it requires the loss of some true planet candidates, which is more difficult in *K2* than *Kepler* because of the systematic noise introduced by the spacecraft roll and thrusters.

To measure how reliable our pipeline is, we invert the light curves to ensure no true transits exist. The pipeline is then run, and any TCEs found and vetted can be considered true FPs. Of 6145 TCEs found in the light curves, only four were able to deceive *EDI-Vetter*. Using Equation (30), we found *EDI-Vetter* to be 94.2% reliable for our catalog. This means that statistically we should only expect four planets within our catalog to be systematic FPs. However, such tests will only address the possibility of systematic FPs. Astrophysical FPs, such as eclipsing binaries, can dilute the transit depth and lead to signal misclassification (Fulton et al. 2017; Matson et al. 2018). Additionally, binary stars can cause inaccurate measurements of planet radius. Ciardi et al. (2017) showed that stellar multiplicity can lead to an overestimation in the number of Earth-sized planets by 15%–20%.

We also measured the completeness of our sample, and find a maximum completeness of 70% for TCEs and 60% for vetted candidates. By injecting artificial transits at the light-curve level of the raw *K2* data, we find that our vetting software is only reducing the recovery rate by about 10%. This is comparable to the loss introduced by *Robovetter* (see Figure 7), indicating that our vetting software is not contributing to the majority of the recovery loss. The lack of completeness is in large part due to the systematic issues of *K2*, increasing the noise and the likelihood of a transit being artificially removed (via detrending) from the light curve.

To ensure that studies of occurrence can be carried out with this catalog, we provide CDPP and window function measurements for each stellar target. These data are available in a machine-readable version, see Table 1.

9.1. Recommendations for Studies of Occurrence

For studies of occurrence we recommend the following procedures.

1. Use a stellar sample where all light curves have a CDPP measure of ≤ 1200 ppm for a transit window of 8 hr. Light curves with more noise are problematic and unlikely to yield usable transits.
2. Only use a target light curve if the window function has a probability > 0.8 for periods of 18 days. Targets with lower probabilities are missing significant portions of the light curve, making transit detections difficult.
3. Consider the issue of systematic correction for radius as seen in Figure 14. This effect could alter the inferred radius distribution of this population of planets.
4. Impose an upper limit on the populations of planet radius. Our eclipsing binary metric is agnostic to stellar radius in order to eliminate any potential biases that a specific stellar catalog may introduce. Thus, planet candidates with a large radius ($R > 20 R_{\oplus}$) may be astrophysical FPs that slipped through our vetting metrics (further discussion on this topic in Section 8).

5. The measure of completeness provided in this study assumes that all planets meet the criterion presented in Equation (11). It is possible that real transiting planets that graze the limb of a star may not meet this criterion. To account for this possibility, we suggest that future studies of occurrence use the following probability function to account for this issue:

$$P\left(\text{detection} \mid \frac{R_{\text{pl}}}{R_{\star}}\right) = \begin{cases} 1; & R_{\text{pl}}/R_{\star} \leq 0.04 \\ 1.04 - R_{\text{pl}}/R_{\star}; & 0.04 < R_{\text{pl}}/R_{\star} \leq 0.3 \\ 0; & 0.3 < R_{\text{pl}}/R_{\star} \end{cases} \quad (35)$$

where P is the probability of detection, assuming a uniform distribution of impact parameters from 0 to 1. This correction is minimal for small measures of the ratio of planet to star radius (R_{pl}/R_{\star}). However, in cases where R_{pl}/R_{\star} is larger, usually in planets orbiting M dwarfs, this probability can get as low as 74%.

We immensely thank the anonymous referee for the careful reading and very thorough comments.

Additionally, we would like to thank Christina Hedges, Geert Barentsen, and Jessie Dotson at the *Kepler* Guest Observer office for helpful discussions about *K2* and this work.

Trevor David provided helpful discussion of *K2* light-curve detrending and transit injections.

Additional special thanks to Megan Bedell for providing us with *Gaia-Kepler* cross-matching out to radius $20''$ upon special request. Also we would like to thank Melanie Swain from the ExoFOP¹⁷ team for helping provide a massive export of stellar data upon request.

This work made use of the *gaia-kepler.fun* cross-match database created by Megan Bedell. The simulations described here were performed on the UCLA Hoffman2 shared computing cluster and using the resources provided by the Bhaumik Institute. This research has made use of the NASA Exoplanet Archive and the Exoplanet Follow-up Observation Program website, which are operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. This paper includes data collected by the *Kepler* mission and obtained from the MAST data archive at the Space Telescope Science Institute (STScI). Funding for the *Kepler* mission is provided by the NASA Science Mission Directorate. STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 526555.

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

Funding for this project was provided by the IPAC Visiting Graduate Fellowship. J.Z. acknowledges funding from NASA ADAP grant 80NSSC18K0431.

Software: EVEREST (Luger et al. 2016, 2018), TERRA (Petigura et al. 2013), K2SFF (Vanderburg & Johnson 2014), K2PHOT (Petigura et al. 2013), PyMC3 (Salvatier et al. 2015), Exoplanet (Foreman-Mackey et al. 2019), RoboVetter

¹⁷ <https://exofop.ipac.caltech.edu>

(Thompson et al. 2018), batman Kreidberg (2015), emcee (Foreman-Mackey et al. 2013).

ORCID iDs

Jon K. Zink  <https://orcid.org/0000-0003-1848-2063>
 Kevin K. Hardegree-Ullman  <https://orcid.org/0000-0003-3702-0382>
 Jessie L. Christiansen  <https://orcid.org/0000-0002-8035-4778>
 Courtney D. Dressing  <https://orcid.org/0000-0001-8189-0233>
 Ian J. M. Crossfield  <https://orcid.org/0000-0002-1835-1891>
 Erik A. Petigura  <https://orcid.org/0000-0003-0967-2893>
 Joshua E. Schlieder  <https://orcid.org/0000-0001-5347-7062>
 David R. Ciardi  <https://orcid.org/0000-0002-5741-3047>

References

- Adams, E. R., Jackson, B., & Endl, M. 2016, *AJ*, 152, 47
 Ballard, S., Christiansen, J. L., Charbonneau, D., et al. 2010, *ApJ*, 716, 1047
 Barros, S. C. C., Demangeon, O., & Deleuil, M. 2016, *A&A*, 594, A100
 Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, *ApJS*, 204, 24
 Berger, T. A., Huber, D., Gaidos, E., & van Saders, J. L. 2018, *ApJ*, 866, 99
 Borucki, W. J. 2016, *RPPh*, 79, 036901
 Bryson, S., Coughlin, J., Batalha, N. M., et al. 2019, arXiv:1906.03575
 Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al. 2010, *ApJL*, 713, L97
 Burke, C. J., & Catanzarite, J. 2017a, Planet Detection Metrics: Window and One-Sigma Depth Functions for Data Release 25, Kepler Science Document, [KSCI-19101-002](https://arxiv.org/abs/19101-002)
 Burke, C. J., & Catanzarite, J. 2017b, Planet Detection Metrics: Per-Target Flux-Level Transit Injection Tests of TPS for Data Release 25, Kepler Science Document, [KSCI-19101-002](https://arxiv.org/abs/19101-002)
 Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, *ApJ*, 809, 8
 Christiansen, J. L. 2017, Planet Detection Metrics: Pixel-Level Transit Injection Tests of Pipeline Detection Efficiency for Data Release 25, Kepler Science Document, [KSCI-19110-001](https://arxiv.org/abs/19110-001)
 Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, *ApJS*, 207, 35
 Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2015, *ApJ*, 810, 95
 Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, *PASP*, 124, 1279
 Ciardi, D. R., Beichman, C. A., Horch, E. P., & Howell, S. B. 2017, *ApJ*, 834, 96
 Claret, A., & Bloemen, S. 2011, *A&A*, 529, A75
 Cloutier, R., & Menou, K. 2019, arXiv:1912.02170
 Coughlin, J. L. 2017, Planet Detection Metrics: Robovetter Completeness and Effectiveness for Data Release 25, Kepler Science Document, [KSCI-19114-002](https://arxiv.org/abs/19114-002)
 Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, *ApJS*, 224, 12
 Crossfield, I. J. M., Ciardi, D. R., Petigura, E. A., et al. 2016, *ApJS*, 226, 7
 Dattilo, Anne, Vanderburg, Andrew, Shallue, Christopher J., et al. 2019, *AJ*, 157, 169
 Dong, S., & Zhu, Z. 2013, *ApJ*, 778, 53
 Dotson, J. L., Barentsen, G., Hedges, C., & Coughlin, J. L. 2019, *RNAAS*, 3, 23
 Dressing, C. D., & Charbonneau, D. 2013, *ApJ*, 767, 95
 Dressing, C. D., & Charbonneau, D. 2015, *ApJ*, 807, 45
 Dressing, C. D., Vanderburg, A., Schlieder, J. E., et al. 2017, *AJ*, 154, 207
 Foreman-Mackey, D., Barentsen, G., & Barclay, T. 2019, dfm/exoplanet: exoplanet v0.1.5, Zenodo, doi:10.5281/zenodo.2587222
 Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
 Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, *AJ*, 154, 109
 Furlan, E., Ciardi, D. R., Everett, M. E., et al. 2017, *AJ*, 153, 201
 Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, *A&A*, 616, A1
 Goodman, J., & Weare, J. 2010, *CAMCS*, 5, 65
 Hardegree-Ullman, K. K., Cushing, M. C., Muirhead, P. S., & Christiansen, J. L. 2019, *AJ*, 158, 75
 Hardegree-Ullman, K. K., Zink, J. K., Christiansen, J. L., et al. 2020, ApJS, in press (arXiv:2001.11511)
 Hippke, M., David, T. J., Mulders, G. D., & Heller, R. 2019, *AJ*, 158, 143
 Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, *ApJS*, 201, 15
 Howell, S. B., Soback, C., Haas, M., et al. 2014, *PASP*, 126, 398
 Hsu, D. C., Ford, E. B., Ragozzine, D., & Ashby, K. 2019, *AJ*, 158, 109
 Huber, D., Bryson, S. T., Haas, M. R., et al. 2016, *ApJS*, 224, 2
 Jenkins, J. M. 2002, *ApJ*, 575, 493
 Jenkins, J. M., Caldwell, D. A., & Borucki, W. J. 2002, *ApJ*, 564, 495
 Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al. 2010, *ApJL*, 713, L87
 Koch, D. G., Borucki, W. J., Rowe, J. F., et al. 2010, *ApJL*, 713, L131
 Kostov, V. B., Mullally, S. E., Quintana, E. V., et al. 2019, *AJ*, 157, 124
 Kreidberg, L. 2015, *PASP*, 127, 1161
 Kruse, E., Agol, E., Luger, R., & Foreman-Mackey, D. 2019, *ApJS*, 244, 11
 Lissauer, J. J., Marcy, G. W., Bryson, S. T., et al. 2014, *ApJ*, 784, 44
 Livingstone, J. H., Endl, M., Dai, F., et al. 2018, *AJ*, 156, 78
 Luger, R., Agol, E., Kruse, E., et al. 2016, *AJ*, 152, 100
 Luger, R., Kruse, E., Foreman-Mackey, D., Agol, E., & Saunders, N. 2018, *AJ*, 156, 99
 Mandel, K., & Agol, E. 2002, *ApJL*, 580, L171
 Matson, R. A., Howell, S. B., Horch, E. P., & Everett, M. E. 2018, *AJ*, 156, 31
 Mayo, A. W., Vanderburg, A., Latham, D. W., et al. 2018, *AJ*, 155, 136
 McCauliff, S. D., Jenkins, J. M., Catanzarite, J., et al. 2015, *ApJ*, 806, 6
 McQuillan, A., Mazeh, T., & Aigrain, S. 2013, *ApJL*, 775, L11
 Mulders, G. D., Pascucci, I., & Apai, D. 2015, *ApJ*, 814, 130
 Mulders, G. D., Pascucci, I., Apai, D., & Ciesla, F. J. 2018, *AJ*, 156, 24
 Mullally, F. 2017, Planet Detection Metrics: Automatic Detection of Background Objects Using the Centroid Robovetter, Kepler Science Document, [KSCI-19115-001](https://arxiv.org/abs/19115-001)
 Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, *ApJS*, 217, 31
 Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2016, *PASP*, 128, 074502
 Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Scikit-learn: Machine Learning in Python (arXiv:1201.0490)
 Petigura, E. A., Crossfield, I. J. M., Isaacson, H., et al. 2018, *AJ*, 155, 21
 Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, *PNAS*, 110, 19273
 Petigura, E. A., Schlieder, J. E., Crossfield, I. J. M., et al. 2015, *ApJ*, 811, 102
 Pope, B. J. S., Parviainen, H., & Aigrain, S. 2016, *MNRAS*, 461, 3399
 Rizzuto, A. C., Vanderburg, A., Mann, A. W., et al. 2018, *AJ*, 156, 195
 Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, *ApJS*, 217, 16
 Salvatier, J., Wiecki, T., & Fonnesbeck, C. 2015, arXiv:1507.08050
 Sanchis-Ojeda, R., Rappaport, S., Winn, J. N., et al. 2014, *ApJ*, 783, 64
 Su, D.-Q., & Cui, X.-Q. 2004, *ChJAA*, 4, 1
 Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, *ApJS*, 235, 38
 Van Cleve, J. E., Howell, S. B., Smith, J. C., et al. 2016, *PASP*, 128, 075002
 Vanderburg, A., & Johnson, J. A. 2014, *PASP*, 126, 948
 Vanderburg, A., Latham, D. W., Buchhave, L. A., et al. 2016, *ApJS*, 222, 14
 von Braun, K., Kane, S. R., & Ciardi, D. R. 2009, *ApJ*, 702, 779
 Youdin, A. N. 2011, *ApJ*, 742, 38
 Yu, L., Crossfield, I. J. M., Schlieder, J. E., et al. 2018, *AJ*, 156, 22
 Ziegler, C., Law, N. M., Baranec, C., et al. 2018, *AJ*, 156, 259
 Zink, J. K. 2019, EDI-Vetter, v1.0, Zenodo, doi:10.5281/zenodo.3585940
 Zink, J. K., Christiansen, J. L., & Hansen, B. M. S. 2019a, *MNRAS*, 483, 4479
 Zink, J. K., & Hansen, B. M. S. 2019, *MNRAS*, 487, 246
 Zink, J. K., Hardegree-Ullman, K. K., Christiansen, J. L., et al. 2019b, *RNAAS*, 3, 43