Planet Detection Metrics:

Statistical Bootstrap Test

KSCI-19086-004

Jon M. Jenkins and Christopher J. Burke

27 July 2016

NASA Ames Research Center
Moffett Field, CA 94035

Prepared by: _____ Date 7-27-16
Jon M. Jenkins, Analysis Lead
(Co-I for Data Analysis)

Concurred by: _____ Date 7/27/16
Christopher J. Burke, Science Office

Approved by: Michael R. Haas Date 7/27/16
Michael R. Haas, Science Office Director

Approved by: _____ Date 7/27/16
Natalie M. Batalha, Project Scientist

Document Control

Ownership

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

Control Level

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

Physical Location

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

Distribution Requests

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

Michael R. Haas
Kepler Science Office Director
MS 244-30
NASA Ames Research Center
Moffett Field, CA 94035-1000

or

Michael.R.Haas@nasa.gov

DOCUMENT CHANGE LOG

CHANGE DATE	PAGES AFFECTED	CHANGES/NOTES
May 15, 2015	All	First issue
August 20, 2015	Pages 10 & 11	Changed number of values reported/plotted
August 26, 2015	Page 10 & 11	Added a paragraph to explain negative boot_mesthresh and boot_mesmean values
July 27, 2016	All	Rewrote to cover 9.1, 9.2 and 9.3 results

Table of Contents

1. Introduction

This document describes the data produced by the Statistical Bootstrap Test over the final three Threshold Crossing Event (TCE) deliveries to NExScI: SOC 9.1 (Q1–Q16)[1] (Tenenbaum et al. 2014), SOC 9.2 (Q1–Q17) aka DR24[2] (Seader et al. 2015), and SOC 9.3 (Q1–Q17) aka DR25[3] (Twicken et al. 2016). The last few years have seen significant improvements in the SOC science data processing pipeline, leading to higher quality light curves and more sensitive transit searches. The statistical bootstrap analysis results presented here and the numerical results archived at NASA's Exoplanet Science Institute (NExScI) bear witness to these software improvements. This document attempts to introduce and describe the main features and differences between these three data sets as a consequence of the software changes.

We first describe the theory underlying the statistical bootstrap test, and then discuss, compare and contrast the results for all three TCE data sets. There are four quantities computed for each Threshold Crossing Event (TCE) produced by the *Kepler* pipeline and they are available at the NExScI Exoplanet Archive[4] as four separate columns in the TCE table: boot_fap, boot_mesthresh, boot_mesmean, and boot_messtd (see Section 4).

Note that the SOC 9.2 Q1–Q17 DR24 bootstrap results are being redelivered in order to improve precision as well as for consistency with the other data sets. All three data sets now use all available data for the bootstrap analysis, rather than just the quarters containing transit features, as was the case for the previous version of the SOC 9.2 results. Using all available data improves the precision of the bootstrap results for TCEs with periods longer than ~93 days as the empirical distributions of null statistics thus obtained can have many more samples, especially for the longest orbital periods. In addition, all three sets of bootstrap results have been produced using the SOC 9.3 algorithm and parameter settings.

We begin with an overview of the Transiting Planet Search (TPS) module of the *Kepler* data processing pipeline in Section 2 and then provide the mathematical development of the statistical bootstrap algorithm in Section 3. The catalog entries for the bootstrap analysis results are defined in Section 4. The bootstrap analysis results are described for the SOC 9.1, SOC 9.2 and SOC 9.3 data sets in Section 5. Section 6 investigates the precision of the bootstrap results as a function of the transit duration and number of transits. We provide detailed results for a sample TCE in section 7.

---

[1] http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q16_tce

[2] http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr24_tce

[3] http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr25_tce

[4] http://exoplanetarchive.ipac.caltech.edu/index.html

2. Overview of Transiting Planet Search

To search for transit signatures, TPS employs a bank of wavelet-based matched filters that form a grid on a three-dimensional parameter space of transit duration, orbital period, and phase (Jenkins 2002; Jenkins et al. 2010). TPS dynamically characterizes the observation noise underlying each flux time series and correlates the reference transit pulse template with the flux time series, accounting explicitly for the correlation structure of the observation noise. This process yields a single event statistic time series, which consists of a component that measures the degree to which the reference transit pulse is correlated with the data, and a component that represents the expected signal-to-noise ratio (SNR) of the reference transit pulse. The single event statistics are folded at each trial orbital period and thresholded to identify statistically significant, transit-like features or threshold-crossing events (TCEs), which are then further scrutinized by the Data Validation (DV) module. DV subjects each TCE to a suite of diagnostic tests to establish or break confidence in the planetary nature of the signature, and fits a limb-darkened transit model to each detected transit-like feature (Wu et al. 2010, Tenenbaum et al. 2010). The threshold, $\eta = 7.1\sigma$, is designed to control the false alarm rate to no more than $\sim 6.24\times10^{-13}$ and thereby restrict the number of statistical false alarms to no more than one over the entire mission, assuming the observation noise is well modeled as Gaussian, though not necessarily white or stationary noise (Jenkins, Caldwell and Borucki 2002).

Since TPS searches each light curve by folding the single-transit detection statistics at each trial orbital period, the detection statistic is referred to as a multiple event statistic (*MES*). Detections in TPS are made under the assumption that the pre-whitening filter applied to the light curve yields a time series whose underlying noise process is stationary, white, Gaussian, and uncorrelated. When the pre-whitened noise deviates from these assumptions, the detection thresholds are invalid and the false alarm probability associated with such a detection may be significantly higher than that for a signal embedded in true, white Gaussian noise (WGN). The Statistical Bootstrap Test, or the Bootstrap, is a way of building the distribution of the null statistics from the data so that the false alarm probability can be estimated for each TCE based on the observed distribution of the out-of-transit statistics using a statistical approach introduced by Ephron (1979). Additionally, the threshold (nominally 7.1 σ) required to control the false alarm rate to $\sim 6.24\times10^{-13}$ can be recalibrated based on the actual observations.

To introduce the Statistical Bootstrap Test, consider a TCE exhibiting p transits of a given duration within its light curve. In this analysis, the light curve is viewed as one realization of a stochastic process. Further, consider the collection of all p-transit detection statistics that could be generated if we had access to an infinite number of such realizations. To approximate this distribution we formulate the single event statistic time series for the light curve and exclude points in transit (plus some padding) for the given TCE. Bootstrap statistics are then generated by randomly drawing from the single event statistics p times with replacement to formulate the p-transit statistic (*i.e.,* the *MES*).

Since the single event statistics encapsulate the effects of local correlations in the background noise process on the detectability of transits, so does each individual bootstrap statistic. So long as the orbital period is sufficiently long (generally longer than several hours), the single event statistics are uncorrelated and the *MES* can be considered as formed by *p* independent random deviates from the distribution of null single event statistics.

Jenkins et al. (2002) formulated a bootstrap test for establishing the confidence level in planetary transit signatures identified in white, but possibly non-Gaussian noise. Jenkins (2002) extended this approach to the case of non-white noise. In both cases, the bootstrap false alarm rate as a function of the *MES* of the detected transit signature was estimated by explicitly generating individual bootstrap statistics directly from the set of out-of-transit data. This direct bootstrap sampling approach can become extremely computationally intensive as the number of transits for a given TCE grows beyond ~15. The number of individual bootstrap statistics that can be formed from the *m* out-of-transit cadences of a light curve and the *p* transits is m^p, which is $\sim 2.9\times10^{48}$ statistics for 10 transits and four years of *Kepler* data. An alternative, computationally efficient method was implemented by formulating the bootstrap distribution in terms of the probability density function (PDF) of the single event detection statistics. The distribution for the *MES* as a function of threshold was then obtained from the distribution of single event statistics.

3. Theoretical Development

In this section we develop the theoretical underpinnings of the Statistical Bootstrap Test, which is based on a mathematical model of the transiting planet search algorithm. Additional background on the historical development of the statistical bootstrap in the context of transiting planet searches can be found in Jenkins, Caldwell and Borucki (2002), Jenkins (2002), and Seader et al. (2015). Here we summarize and expand upon previous work by detailing more recent developments in the implementation of the algorithm.

TPS constructs a (multiple event) detection statistic, *Z,* for each trial transit pulse duration, orbital period and orbital phase. The *MES*, *Z*, can be expressed as

$$Z = \sum_{i\in S} C(i)/\sqrt{\sum_{i\in S} N(i)}, \quad (1)$$

where *S* is the set of transit times that a single period and epoch pair select out, *C(i)* is the correlation time series formed by correlating the whitened data to a whitened transit signal template with a transit centered at the i^{th} timestep in the set *S*, and *N(i)* is the template normalization time series. The square root of the normalization time series,

$\sqrt{N(i)}$, is the expected value of the *MES* or SNR for the reference transit pulse.[5]

If the observation noise process underlying the light curve is well modeled as a Gaussian noise process that is possibly non-white and/or non-stationary, then the single event statistics will be zero-mean, unit-variance Gaussian random deviates. The false alarm rate of the transit detector would then be described by the complementary distribution for a zero-mean, unit-variance Gaussian distribution:

$$\overline{F}_Z(Z) = \frac{1}{2} erfc\left(Z/\sqrt{2}\right), \tag{2}$$

where $erfc(\cdot)$ is the standard complementary error function. If the power spectral density of the noise process is not perfectly captured by the whitener in TPS, then the null statistics will not be zero-mean, unit-variance Gaussian deviates. A bootstrap analysis allows us to obtain a data-driven approximation of the actual distribution of the null statistics, rather than relying on the assumption that the pre-whitener is perfect.

Let Z_p denote a p-transit multiple event statistic. The random variable Z_p is a function of the random variables corresponding to the correlation and normalization terms in the single event statistic time series, $\{C(i), N(i)\}$, and is thus governed by a bivariate distribution with components

$$C_p = \sum_{i \in S} C(i) \text{ and } N_p = \sum_{i \in S} N(i), \tag{3}$$

given the definition of the multiple event statistic in Eq. 1. The joint density of C_p and N_p can be determined from the joint density of the single event statistic components C and N as

$$f_{C_p,N_p}(C_p, N_p) = f_{C,N}(C,N) * f_{C,N}(C,N) * \ldots * f_{C,N}(C,N), \tag{4}$$

where '$*$' is the convolution operator and the convolution is performed p times. This follows from the fact that the bootstrap samples are constructed from *independent* draws from the set of null (single event) statistics with replacement.[6] Given that convolution in the time/spatial domain corresponds to multiplication in the Fourier domain, Equation (4) can be represented in the Fourier domain as

---

[5] The inverse of $\sqrt{N(i)}$ can be interpreted as the effective, white Gaussian noise "seen" by the reference transit and is the definition for the combined differential photometric precision (CDPP) reported for the *Kepler* light curves at 3, 6, and 12 hours duration.

[6] In this implementation, we choose to assume that the null statistics are governed by a single distribution. If this is not the case (for example, if the null statistic densities vary from quarter to quarter), then Equation (4) could be modified to account for the disparity in the relevant single event statistics in a straightforward manner.

$$\Phi_{C_p,N_p} = \Phi_{C,N} \cdot \Phi_{C,N} \cdot \ldots \cdot \Phi_{C,N} = \Phi_{C,N}^{P}, \quad (5)$$

where $\Phi_{C,N} = \Im\{f_{C,N}\}$ is the 2-D Fourier transform of the joint density function $f_{C,N}$. Here, the arguments of the Fourier transforms of the density functions have been suppressed for clarity. The use of 2-D fast Fourier transforms results in a highly tractable algorithm from a computational point of view. Figure 1 illustrates the construction of the *MES* distribution for the case of a 4-transit TCE.

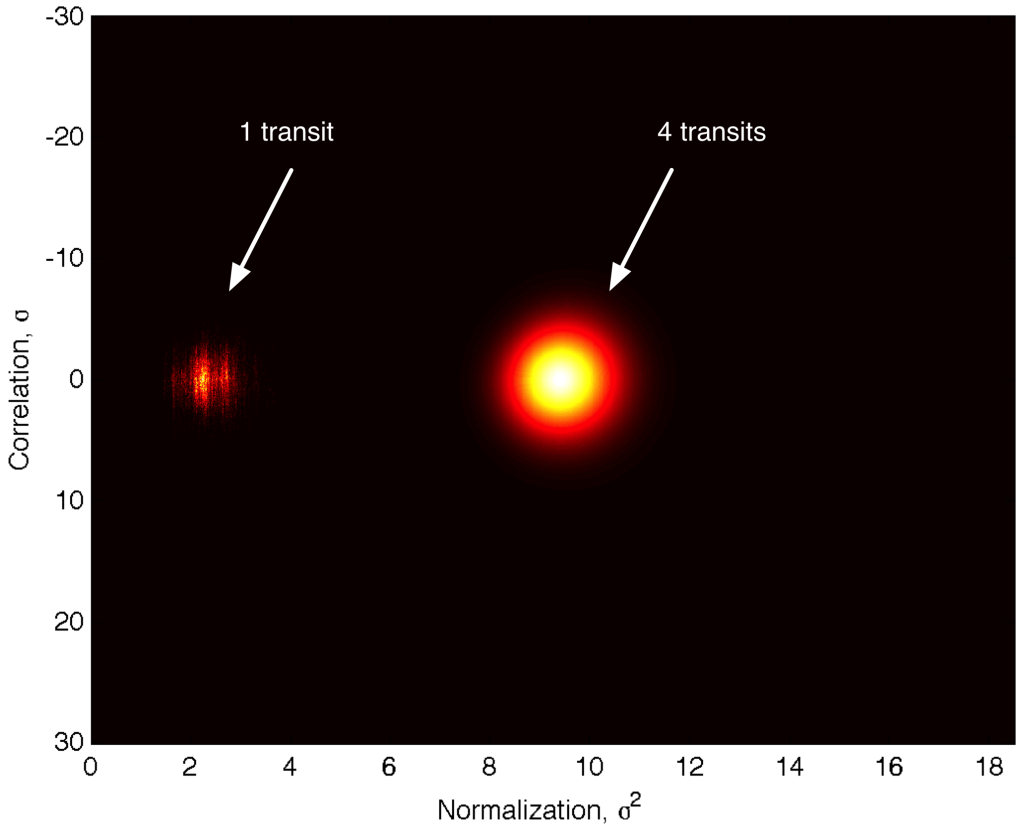

Figure 1. False color image of the correlation component *C* versus the normalization component *N* for the bivariate distributions for the single event null statistics and for the 4-transit multiple event statistic distribution for one TCE. Although the 1-transit distribution is highly irregular, that for the 4-transit distribution is much more symmetric and Gaussian, as follows from the central limit theorem.

The implementation of Equation (5) requires that a 2-D histogram be constructed for the $\{C,N\}$ pairs over the set of single event null statistics. Care must be taken to manage the size of the histogram to avoid spatial aliasing as the use of fast Fourier transforms (FFTs) corresponds to circular convolution. We chose to formulate the 2-D grid to allow for as high as $p = 8$ transits in order to control the memory required for the computations. The intervals covered by the realizations (*i.e.,* the support) for each of C and N were sampled with 256 bins, and centered in a 4096 by 4096 array. When $p > 8$, it is necessary to implement Equation (5) iteratively in stages after each of which the characteristic

function is transformed back into the spatial domain, then bin-averaged by a factor of 2, and padded back out to the original array size. Care must also be taken to manage knowledge of the zero-point of the histogram in light of the circular convolution, as it shifts by 1/2 sample with each convolution operation in each dimension.

Once the *p*-transit 2-D density function $f_{C_p,N_p}(C_p,N_p)$ is obtained, it can be "collapsed" into the sought-after 1-D density, $f_Z(Z)$, by mapping the sample density for each cell with center coordinates $\{C_i,N_j\}$ to the corresponding coordinate $Z_{i,j} = C_i/\sqrt{N_j}$, and formulating a histogram with a resolution of, say, 0.1σ in *Z* by summing the resulting densities that map into the same bins in *Z*. Due to the use of FFTs, the precision of the resulting density function is limited to the floating-point precision of the variables and computations, which is ~2.2×10^{-16} for double precision arithmetic. For small *p*, the density may not reach the limiting numerical precision because of small number statistics, and for large *p*, round-off errors can accumulate below about 10^{-14}. However, the bootstrap results can be extrapolated to high *MES* values by fitting the mean, μ, and standard deviation, σ, of a Gaussian distribution to the upper tail of the empirical distribution in the region $10^{-4} \le \overline{F}_Z \le 10^{-13}$ using the standard complementary error function:

$$\overline{F_Z}(Z) = 0.5 erfc\left((Z-\mu)/\sqrt{2}\sigma\right). \quad (6)$$

Note that the fitted distribution can only be used to extrapolate the upper tail of the bootstrap distribution, and is not valid for describing the core of the empirical distribution, as it is not constrained to fit the latter below 10^{-4}.

In order to simulate the use of χ-square vetoes (Seader et al. 2013) that effectively remove strong transient and impulsive features that trigger TPS but which are inconsistent with physical transit signatures, we pre-filtered the single event statistic time series to remove the three most positive peaks and their "shoulders" down to 2σ. The three most negative peaks were also handled in a similar fashion to avoid biasing the mean of the null statistics in a negative direction. We also identified and removed points with a density of zero-crossings that fell below 1/4 that of the median zero-crossing density. This step removed single event statistics in regions where the correlation term experienced strong excursions from zero due to unmitigated sudden pixel sensitivity dropouts and thermal transients near monthly and quarterly boundaries. Typically, these pre-filters retained more than 99% of the original out-of-transit single event statistics.

4. Column Definitions

There are four quantities derived from the bootstrap test that are present in the TCE Table. These four quantities are defined as follows:

boot_fap: The false alarm probability is the integral of the distribution of the null *MES* statistics above the *MES* of the detection. The distribution of the null *MES* statistics is constructed by the bootstrap test. Nominally, the null *MES* is Gaussian distributed with zero mean and unit variance. In reality however, due to imperfections in the whitening process, uncorrected systematics, etc., the distribution of the null *MES* deviates from this nominal distribution form.

boot_mesthresh: The search threshold required, given the distribution of the null *MES* estimated from the bootstrap algorithm, to achieve the same false alarm probability as that of a 7.1 σ threshold on a Gaussian distribution with zero mean and unit variance ($\bar{F}_Z$~6.24×10^{-13}).

boot_mesmean: The mean of the best-fit Gaussian distribution to the upper tail ($10^{-13} \leq \bar{F}_Z \leq 10^{-4}$) of the null *MES* distribution estimated by the bootstrap. This quantity, together with the quantity boot_messtd, is useful for extrapolating the false alarm probability to values less than 10^{-13} and for high *MES* values.

boot_messtd: The standard deviation of the best-fit Gaussian distribution to the upper tail ($10^{-13} \leq \bar{F}_Z \leq 10^{-4}$) of the null *MES* distribution estimated by the bootstrap.

In cases where there is not enough data to run the bootstrap, the false alarm probability is set to -1. In some cases, the bootstrap test can't use the data from the distribution it has constructed to interpolate for the false alarm probability; rather, it must extrapolate because the *MES* is outside the regime that the distribution covers. To do the extrapolation, a robust fit of an error function is done in log space to the Cumulative Distribution Function (CDF) of the *MES* with $10^{-13} \leq \bar{F}_Z \leq 10^{-4}$. The parameters of the fit are used to calculate the false alarm probability for the *MES* of the detection. Features in the CDF can sometimes cause the fit to be poor, which in turn causes the fit parameters and resulting false alarm probabilities to also be poor.

5. Results

This section describes the results for each TCE data set and compares and contrasts the results across the three data sets.

5.1. Q1–Q16

Figures 2 and 3 show the false alarm probability for 16,014 TCEs as a function of the *MES* value of each TCE. A small fraction of the TCEs' light curves had too few points remaining after removing in-transit and neighboring points to conduct a bootstrap analysis. We matched the Q1–Q16 TCEs against the cumulative KOI table through Data Release 24 (Coughlin et al. 2016) on NExScI's exoplanet archive. The KOIs (consisting of a mixture of planet candidates, confirmed/validated planets and astrophysical false positives) tend to lie in a band whose left edge is approximately enveloped by the curve

expected for zero-mean, unit variance (ZMUV) Gaussian noise. At low SNR, there is a small population of planet candidates and astrophysical false positives with false alarm rates above 10^{-10} that are embedded in a much larger population of TCEs that form a horizontal "cloud" where the false alarm rate is nearly independent of the *MES*. Visual inspection of the light curves underlying points in this "cloud" indicate that most of these light curves are polluted with residual stellar variability and flares. Many appear consistent with being red giants with power spectral densities richly populated by pressure mode oscillations. The whitener in TPS is not designed to handle such noise, resulting in spurious detections.

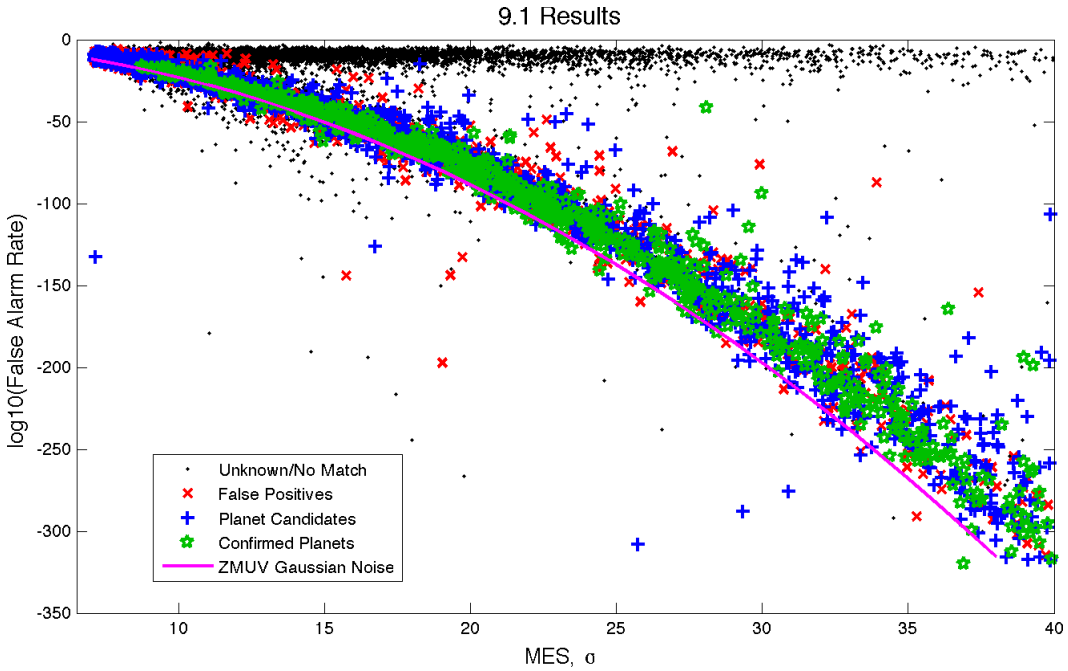

Figure 2. False alarm rate as a function of the multiple event statistic (*MES*) for each of the 16,014 TCEs returning bootstrap results in the Q1–Q16 transiting planet search. The points are colored by the dispositions of the TCEs in NExScI Exoplanet Archive's cumulative KOI table through DR24. The magenta line indicates the expected value for a zero-mean, unit-variance Gaussian process.

Most of the points falling below the ZMUV curve are very short period TCEs with 500 or more transits. These targets have little data left after the removal of in-transit samples and the remaining single event null statistics are slightly biased with a mean below zero. For periods sufficiently short so that there are ~500 or more transits, computing the bootstrap distribution for the *MES* involves raising the 2-D characteristic function for the single event null statistics to the number of observed transits (see Eq. 5). This process yields a distribution with a mean that is significantly negative. This results in false alarm rates well below those expected for ZMUV noise for such cases.

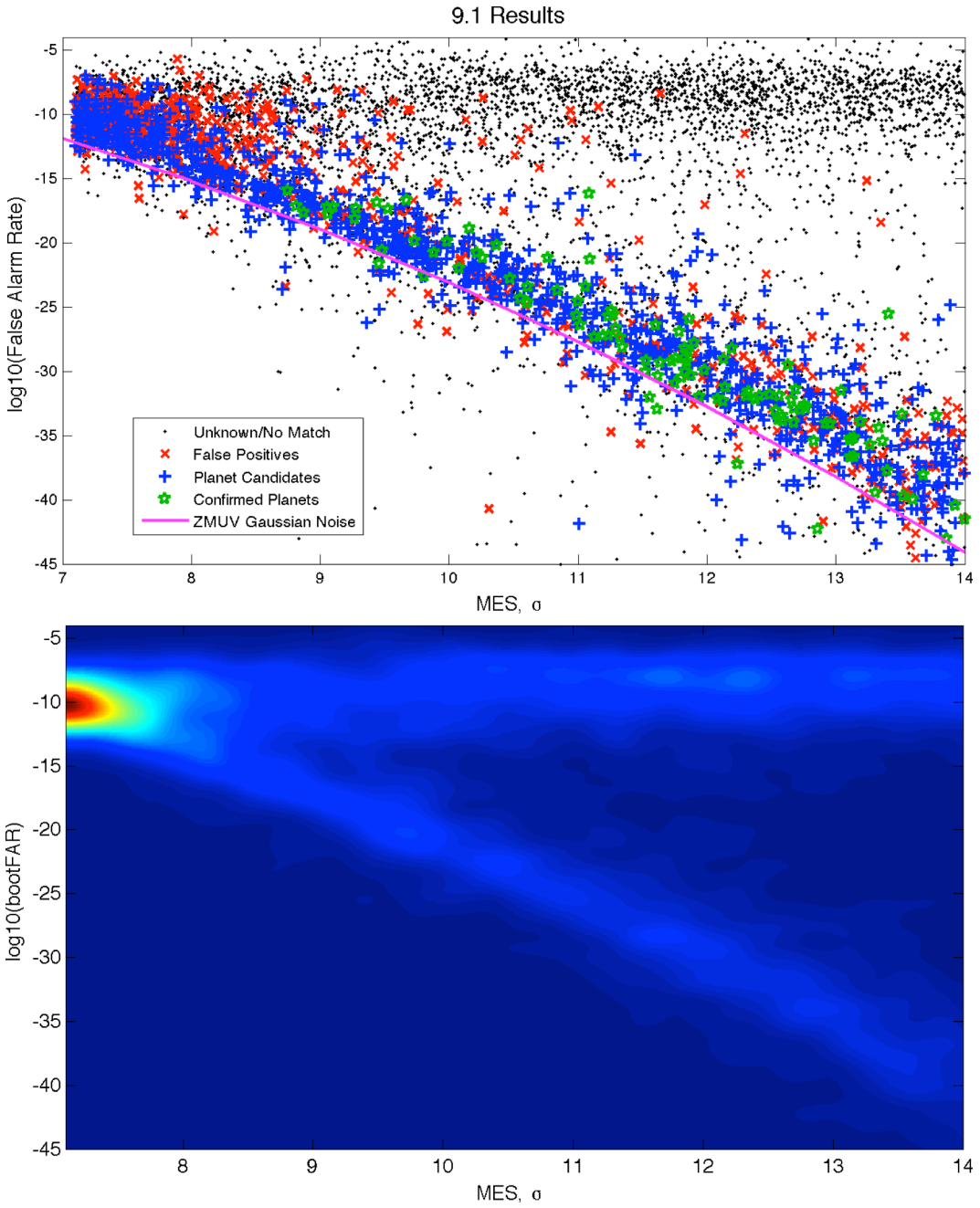

Figure 3. Zoomed view of Figure 2. Top panel: false alarm rate as a function of the *MES* for each of the Q1–Q16 TCEs. Bottom panel: Density plot of the false alarm rate as a function of the *MES*. Note that the two principal populations, the horizontal branch with little dependency on *MES* and the one that is approximately enveloped by the expected curve for ZMUV Gaussian noise both merge at low SNR (<9 σ) near log(FAR) = 10^{-10}.

Figure 4 shows a plot of the false alarm rate as a function of the bootstrap threshold

(boot_mesthresh) for the SOC 9.1 Q1–Q16 TCEs, colored by disposition for the KOIs matched against the TCE ephemerides.[7] Note that some confirmed/validated planets and planet candidates have thresholds above ~10σ. The confirmed planets include Kepler-90d and h, Kepler-30c and d, and Kepler-444e and f. All of these systems exhibit strong transit timing variations and therefore the use of a linear ephemeris to mask out transits leaves residual transits in the null statistic time series, inflating the bootstrap thresholds and false alarm rate estimates for these cases.

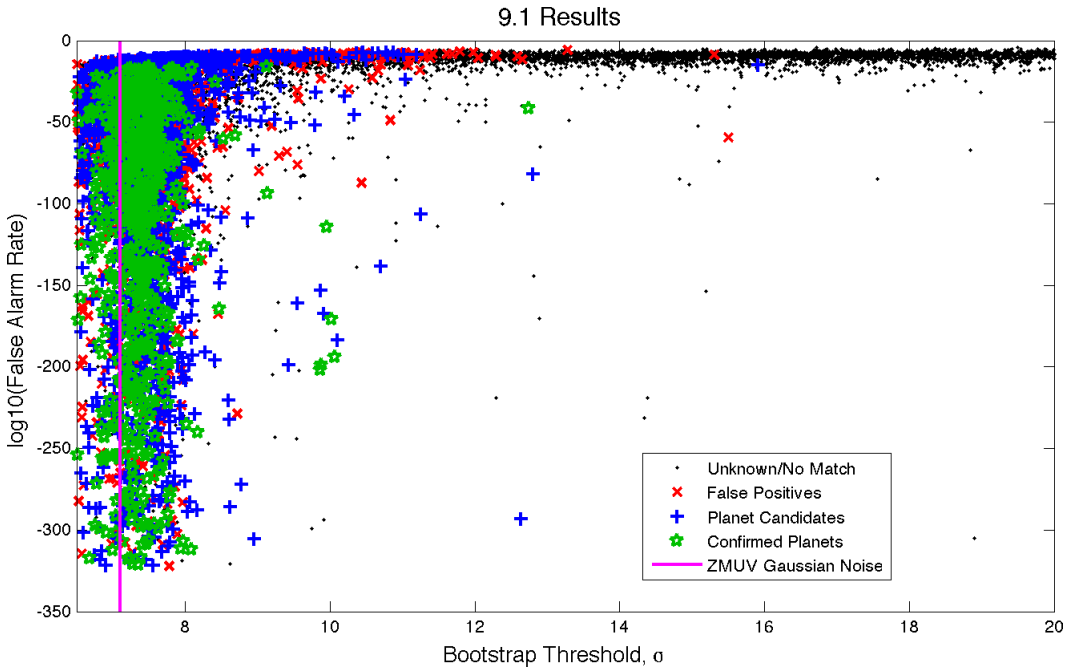

Figure 4. False alarm rate as a function of the bootstrap threshold for the Q1–Q16 TCEs. For the KOIs recovered in this data set, the points are colored by the disposition in the cumulative DR24 KOI table. The vertical line at 7.1σ represents the threshold expected for ZMUV Gaussian noise.

We note that the bootstrap threshold (boot_mesthresh) is negative for a few stars. This is most often the case for very short orbital periods (≲5 days) where the residual single event statistics are so cut up by the removal of transit signatures that they are biased negative relative to the true mean and the resulting N-transit statistics are significantly skewed. This is a limitation of the methodology, but is expected behavior. The boot_mesmean can be negative as well because the upper tail of the empirical bootstrap distribution is being modeled and the fit is unconstrained. The model is only useful for extrapolating to *MES* values above that observed in the empirical distribution where the

[7] For this document, a match was considered valid if both the epoch and the period of the TCE were within 0.1 days of the given KOI.

false alarm rate is $\ll 10^{-4}$. If a valid fit cannot be obtained for some reason, then the Gaussian fit parameters are gapped (i.e., marked as unpopulated).

While the bootstrap results for the well-behaved transit signatures and the spurious detections in the horizontal cloud are not well separated at low SNR (<9 σ), the bootstrap false alarm probabilities are strong indicators for high SNR TCEs. Improvements in the quality of the light curves and the transiting planet search codebase dramatically improve the situation, as will be seen in Section 5.3.

5.2. SOC 9.2 Q1–Q17 DR24

The SOC 9.3 bootstrap algorithm was run on the SOC 9.2 Q1–Q17 DR24 TCEs previously and archived at NExScI (see KSCI-19086-003). That set of bootstrap results only included quarters with transits in the calculation. This note documents redelivery of the bootstrap results for this data set using all available data from Q1–Q17 in the bootstrap analysis, for consistency with the other two data sets delivered at this time.

The results are similar to those for the Q1–Q16 data set, although since the SOC 9.2 bootstrap was used as a veto in TPS for this run, the horizontal "cloud" of high *MES*/high false alarm rate objects is missing, as almost all of these objects were rejected in TPS and not presented to DV for further analysis and characterization (see Christiansen et al. 2016 for a detailed discussion of the impact of the bootstrap veto on completeness of the transit search). Figures 5 and 6 present the bootstrap false alarm rate as a function of the *MES*, while Figure 7 presents the bootstrap false alarm rate as a function of the bootstrap threshold.

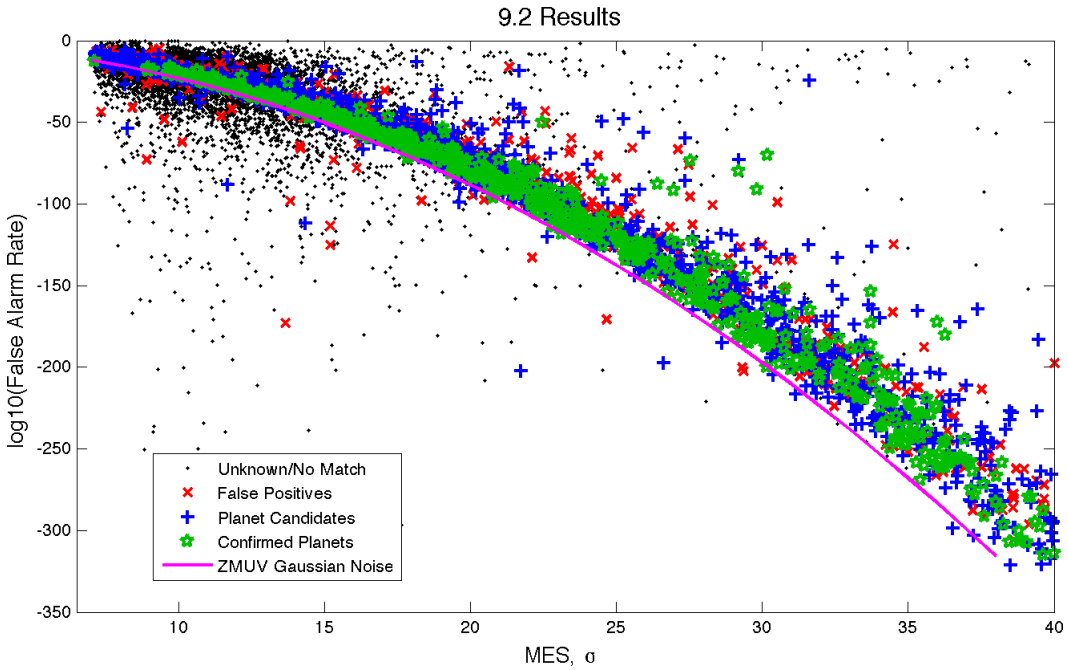

Figure 5. False alarm rate as a function of the *MES* for each of the 19,856 TCEs returning bootstrap results in the Q1–Q17 DR24 transiting planet search. The points are colored by KOI disposition. The magenta line indicates the expected value for a ZMUV Gaussian process.

Because the horizontal "cloud" feature is completely absent from Figure 5, we can conclude that the bootstrap is quite effective at filtering non-transit-like features associated with this population from the TPS results. The KOIs identified in the SOC 9.2 DR24 TCEs fall along a band that is enveloped on the left by that expected for ZMUV Gaussian noise. As with the SOC 9.1 results, none of the confirmed or validated planets

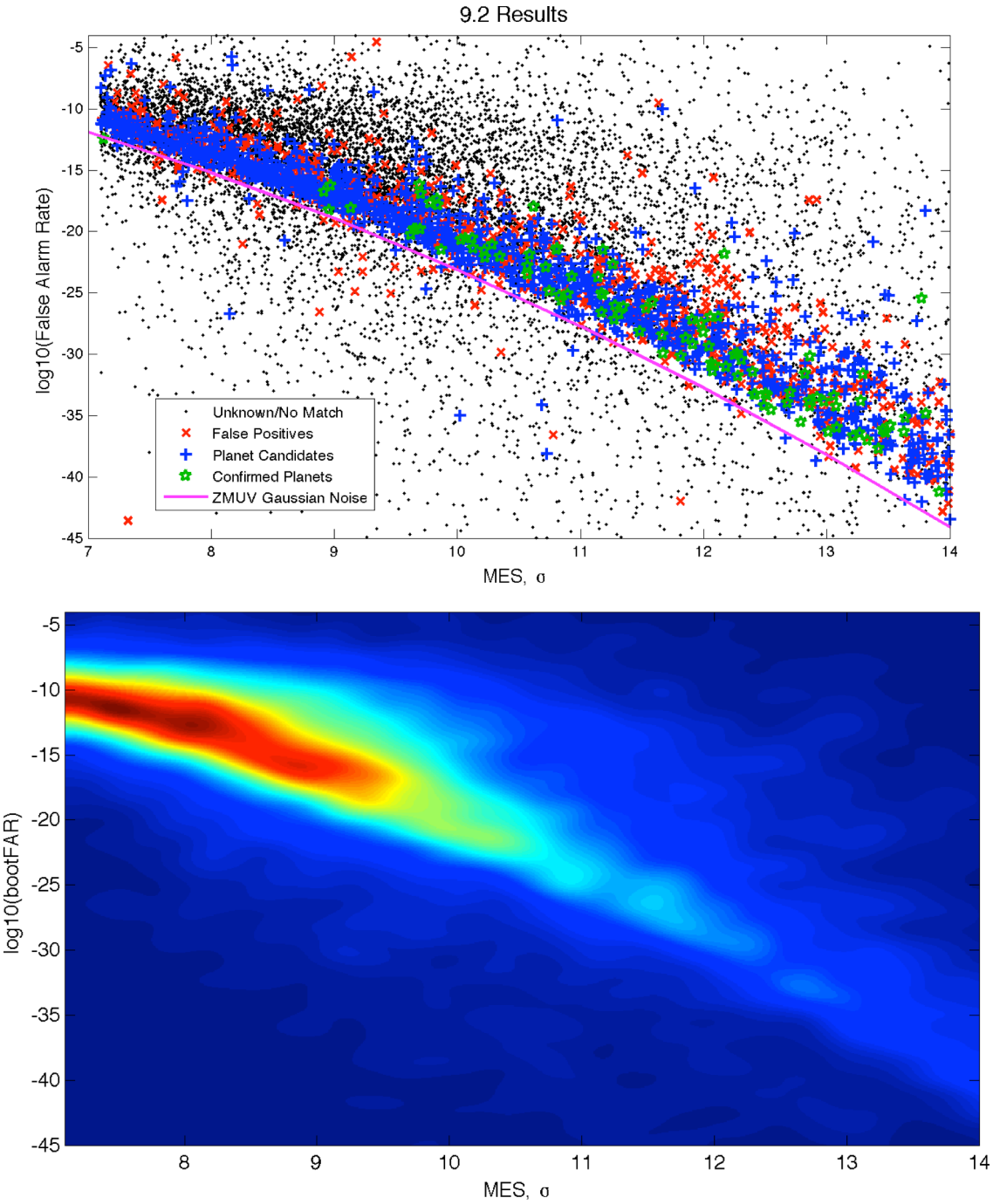

Figure 6. Zoom of Figure 5. Top panel: false alarm rate vs. *MES* for the 19,856 SOC 9.2 Q1–Q17 DR24 TCEs. Bottom panel: Density plot of the false alarm rate as a function of the multiple event statistic. Note that the horizontal branch with little dependency on *MES* is missing for SOC 9.2 as the bootstrap was used as a veto in TPS.

have bootstrap false alarm rates above 10^{-12}, suggesting that the bootstrap can be used to screen against spurious TCEs, although it is unclear whether this can be done without

rejecting true transiting planets at low SNR (<9 σ) given that the large population of spurious TCEs evident for SOC 9.1 are not present in this SOC 9.2 data set.

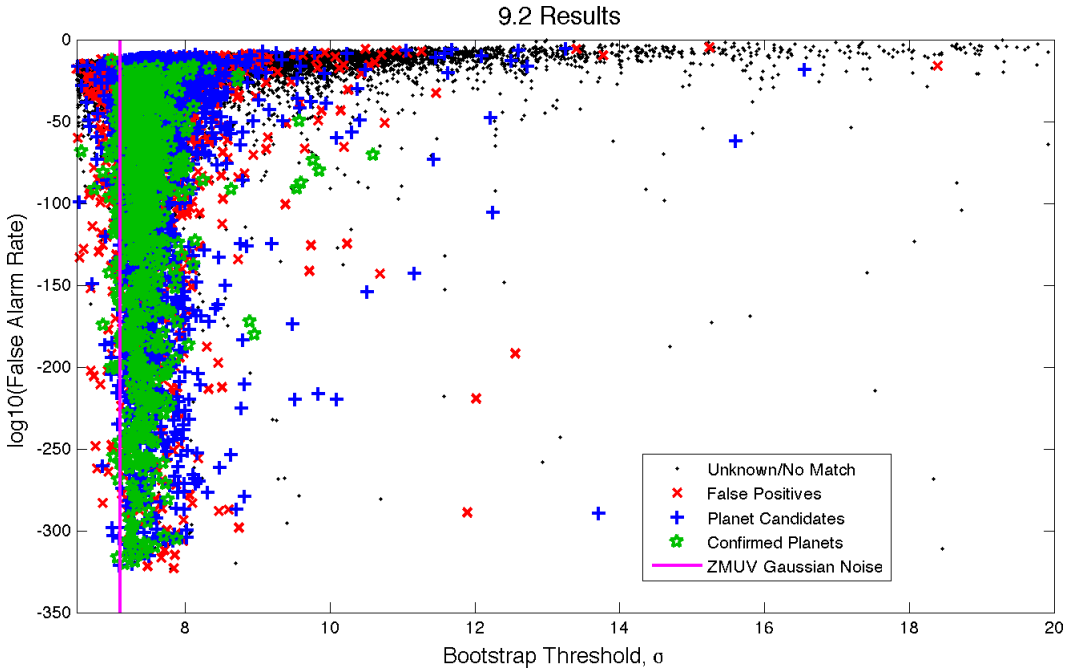

Figure 7. False alarm rate as a function of the bootstrap threshold for the 19,856 Q1–Q17 DR24 TCEs, colored by KOI disposition. The vertical line at 7.1σ represents the threshold expected for ZMUV Gaussian noise.

5.3. SOC 9.3 Q1–Q17 DR25

Figures 8 and 9 present the false alarm rate as a function of *MES* for the 24,179 TCEs returning bootstrap results in the SOC 9.3 Q1–Q17 DR25 search. It is interesting to note that the ZMUV curve is a much better envelope for the band of TCEs that are also KOIs. Note also that this band of KOIs is tighter than those for either of the two previous data sets. This is largely due to changes made in the SOC 9.3 TPS codebase to improve the performance of the whitening filter as well as changes in the assignment of photometric apertures (Smith et al. 2016) and improvements in systematic error correction. These changes and their affect on the DR25 TCEs are documented in Twicken et al. (2016).

For the SOC 9.3 DR25 results, there is much better separation of the low reliability TCEs in the horizontal "cloud" population from those in the band following the ZMUV curve compared to the SOC 9.1 TCEs. The unreliable TCEs have false alarm probabilities greater than approximately 10^{-11} while all confirmed and/or validated planets, almost all planet candidates, and most astrophysical false positives have $\log(\text{FAR}) < 10^{-11}$.

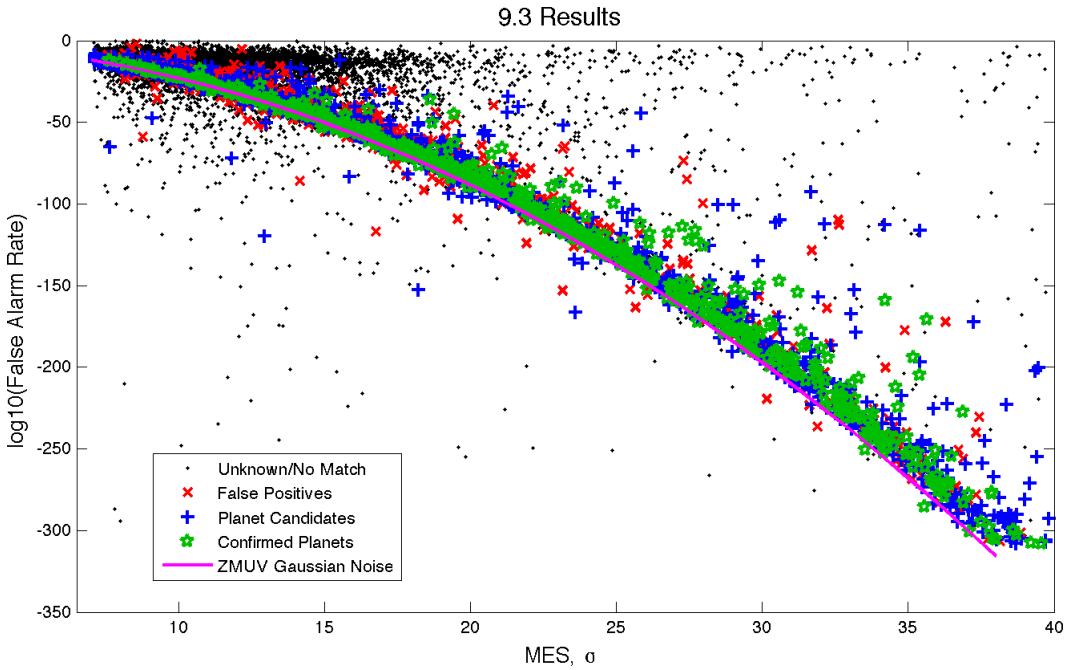

Figure 8. False alarm rate as a function of the multiple event statistic (*MES*) for each of the 24,179 TCEs returning bootstrap results in the SOC 9.3 Q1–Q17 DR25 transiting planet search. The points are colored by the dispositions of the TCEs in NExScI Exoplanet Archive's cumulative DR24 KOI table. The magenta line indicates the expected value for a ZMUV Gaussian process.

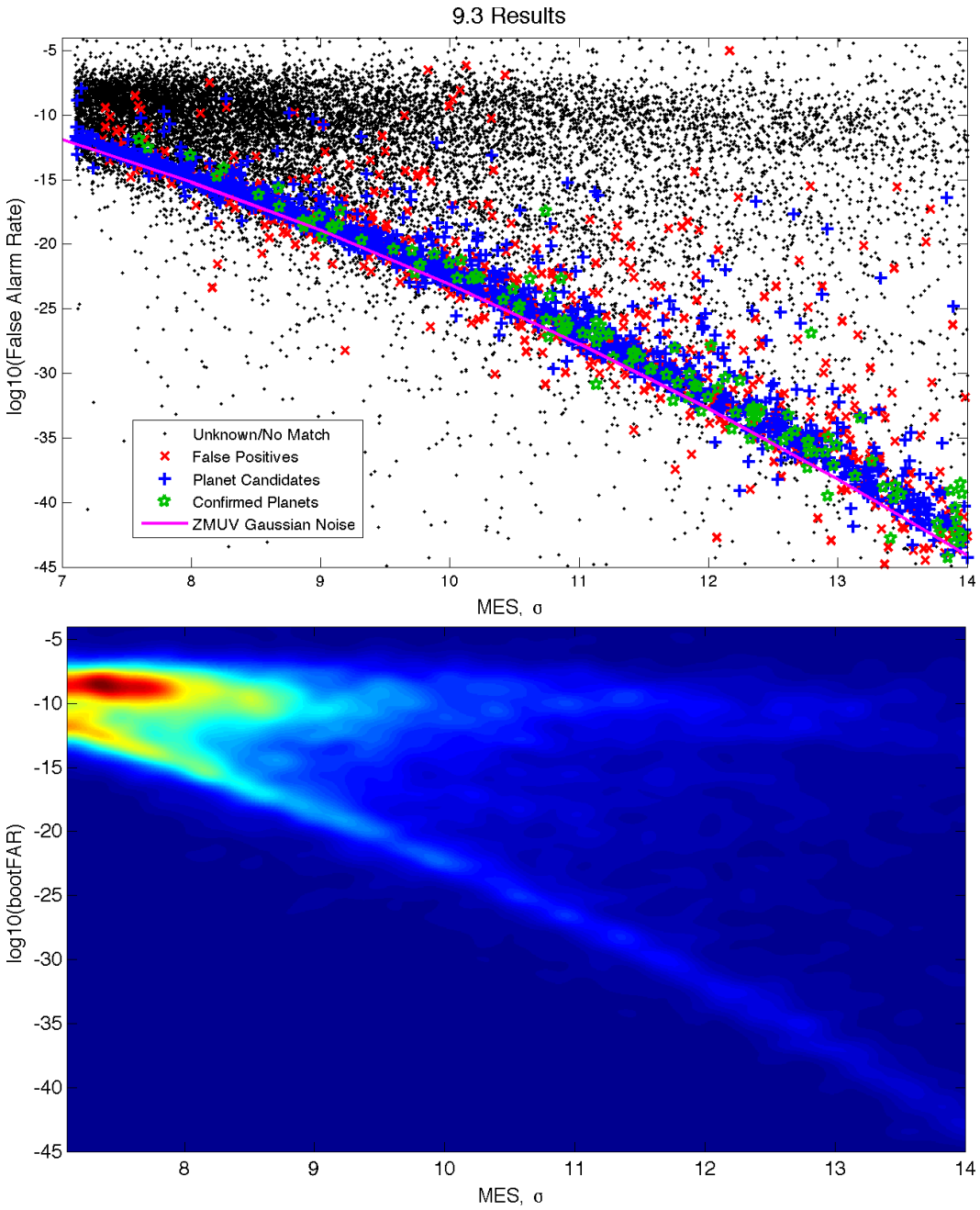

Figure 9. Zoom of Figure 8. Top panel: false alarm rate as a function of the *MES* for each of the SOC 9.3 Q1–Q17 DR25 TCEs, colored by KOI disposition. Note that the "cloud" of high *MES*/high FAR reappears here as the bootstrap was *not* used as a veto in TPS for this run. Bottom panel: Density plot of the false alarm rate as a function of the *MES*. Note that the two principal populations, the horizontal branch with little dependency on *MES* and the one that is approximately enveloped by the expected curve for ZMUV Gaussian noise are well separated in the SOC 9.3 results with a valley between them at log(FAR) ≈ 10^{-11}.

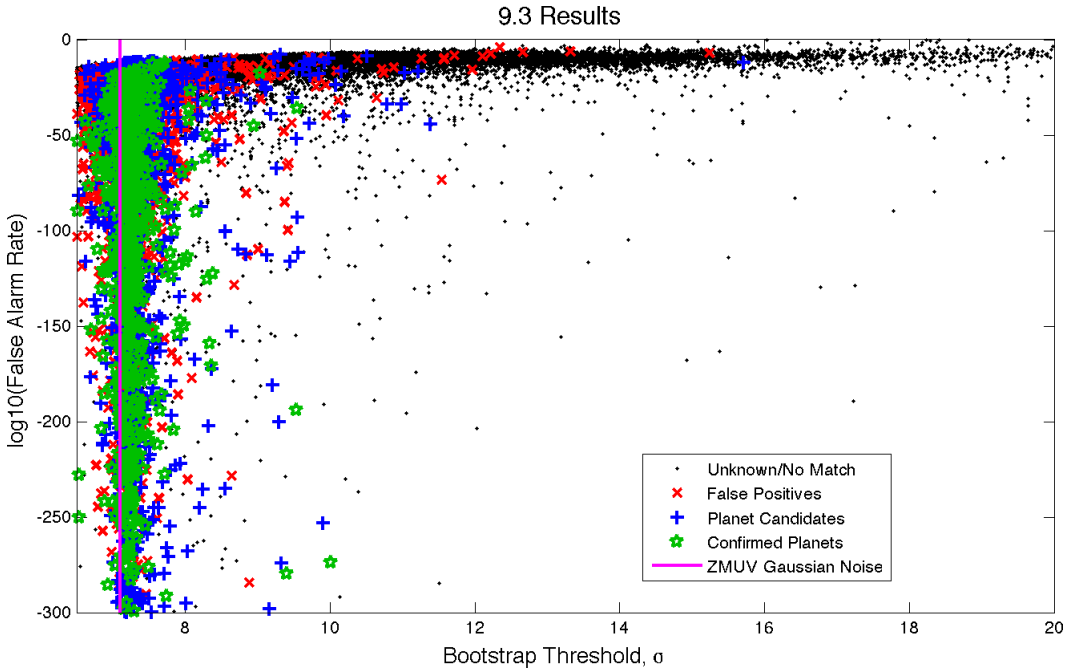

Figure 10. False alarm rate as a function of the multiple event statistic (*MES*) for each of the 24,179 TCEs returning bootstrap results in the Q1–Q17 DR25 transiting planet search, colored by the KOI disposition. The magenta line indicates the expected value for a ZMUV Gaussian process.

Figure 10 shows the bootstrap false alarm rate as function of bootstrap threshold. The SOC 9.3 changes also reduced the number of planet candidates with bootstrap thresholds >10σ from 46 in SOC 9.1 (see Figure 4) and 56 in SOC 9.2 (see Figure 7) to only 13 in SOC 9.3.

5.4. Comparison Across Data Sets

While the results of the bootstrap analysis are similar across the three data sets, there are differences due in small part to the amount of data considered in each one (16 quarters of data vs. 17 quarters of data), and in large part to code base changes for the production of the light curves. In this section we discuss some of the differences between the data sets.

Figure 11 shows the difference in the log of the bootstrap false alarm rates between SOC 9.2 and SOC 9.1, colored by KOI disposition, where the 9.2 false alarm rates are calculated using the SOC 9.1 *MES* values to ensure that the comparisons are valid. For false alarm rates exceeding $\sim 10^{-40}$ there is a relatively tight core of points between $\sim 10^{1.6}$ and $\sim 10^{-0.6}$ for the KOIs. The points are well correlated between the two data sets considering the differences in the code bases and the fact that results smaller than $\sim 10^{-16}$ are all extrapolations from fits to the upper tails of the empirical bootstrap distributions.

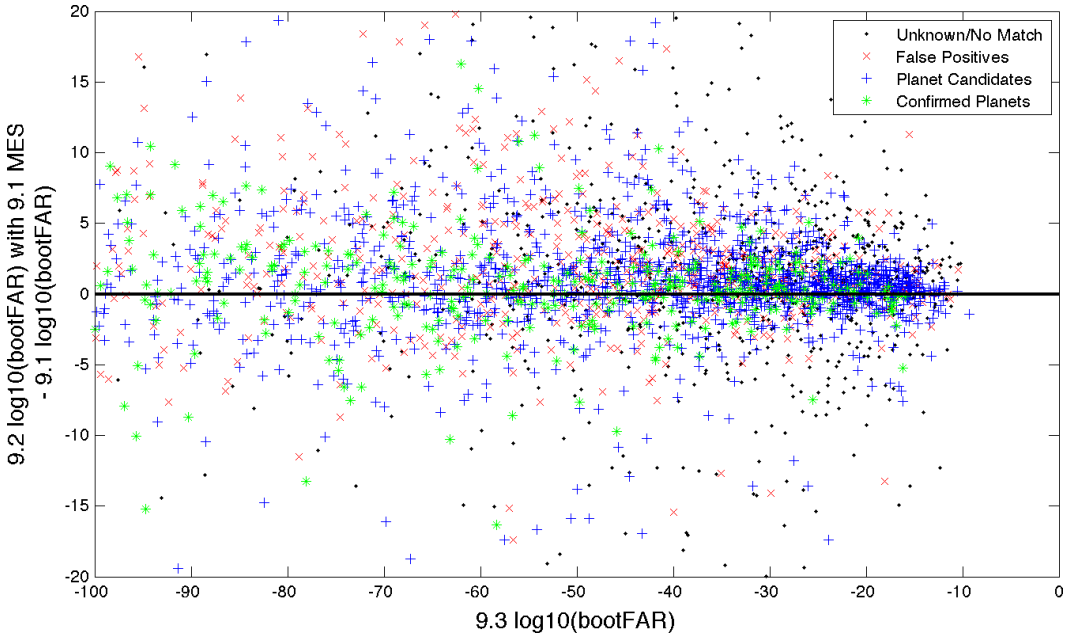

Figure 11. Difference in false alarm rate for the SOC 9.2 Q1–Q17 TCEs vs. that for the SOC 9.1 Q1–Q16 TCEs as a function of the SOC 9.1 false alarm rates for those objects that have matching ephemerides in the two data sets. The SOC 9.2 false alarm rates are calculated for the SOC 9.1 *MES* values to ensure a valid comparison. The SOC 9.2 false alarm rates are slightly higher than those for SOC 9.1, perhaps due to differences in the code bases.

Figure 12 shows the difference in the log of the bootstrap false alarm rates between SOC 9.3 and SOC 9.1, colored by KOI disposition, where the 9.3 false alarm rates are calculated using the SOC 9.1 *MES* values to ensure that the comparisons are valid. For false alarm rates greater than $\sim 10^{-40}$ there is a relatively tight core of points between $\sim 10^{0.5}$ and $\sim 10^{-2}$ for the confirmed/validated planets, planetary candidates and astrophysical false positives. The SOC 9.3 results are somewhat lower than those for the same objects for SOC 9.1, reflecting the improvements in the flux time series and in the TPS algorithm due to the SOC 9.3 codebase changes.

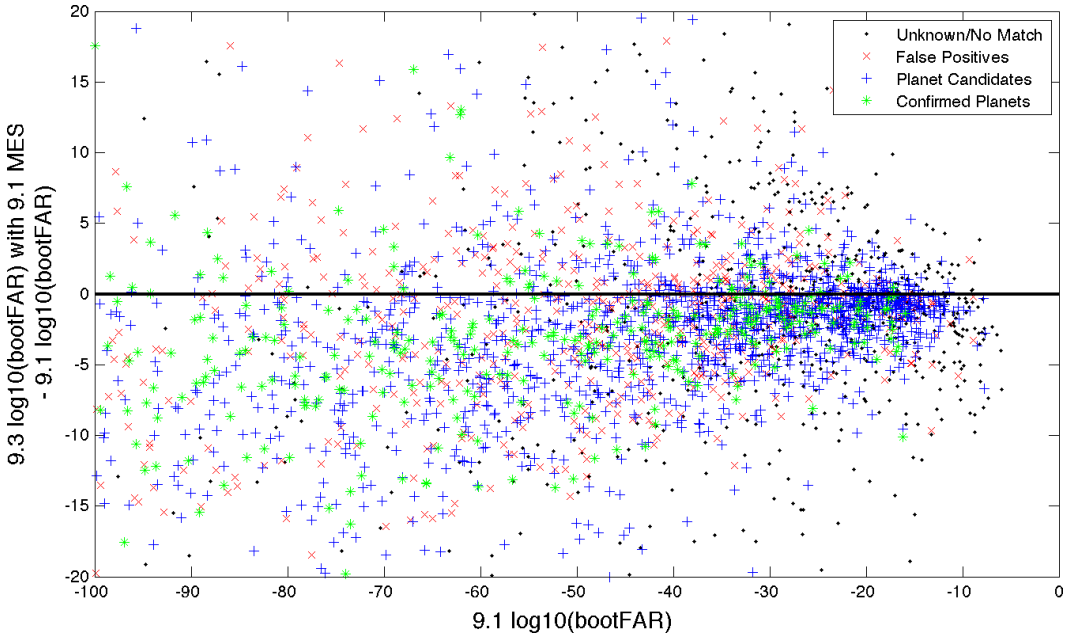

Figure 12. Difference in False alarm rate for the SOC 9.3 Q1–Q17 TCEs vs. that for the SOC 9.1 Q1–Q16 TCEs as a function of the SOC 9.1 false alarm rates for those objects that have matching ephemerides in the two data sets. The SOC 9.3 false alarm rates are calculated for the SOC 9.1 *MES* values to ensure a valid comparison. The SOC 9.3 false alarm rates are somewhat lower than those for SOC 9.1, due to the improvements in the photometric pipeline as well as in TPS.

The codebase changes from SOC 9.1 to SOC 9.3 also tightened the distribution of bootstrap thresholds produced by the statistical bootstrap analysis. In addition, the median threshold dropped from 7.44 σ in SOC 9.1 to 7.24 σ in SOC 9.3, demonstrating the increase in sensitivity as the SOC codebase evolved. The 10th, 50th and 90th percentiles for the bootstrap thresholds for TCEs that were matched against confirmed/validated planets and astrophysical false positives are given in Table I.

Table I: Summary Statistics of the Bootstrap Thresholds for KOIs

	10th Percentile	50th Percentile	90th Percentile
SOC 9.1	6.84 σ	7.44 σ	8.36 σ
SOC 9.2	7.17 σ	7.51 σ	8.36 σ
SOC 9.3	7.02 σ	7.24 σ	7.75 σ

Figure 13 shows histograms of the bootstrap thresholds for KOIs for each of the data sets. The SOC 9.3 results show a broad upper tail, thanks to improved sensitivity in the search, and relaxation of the vetoes in TPS for the final run, which doubled the number of TCEs from 16,285 in SOC 9.1 to 34,032 in SOC 9.3.

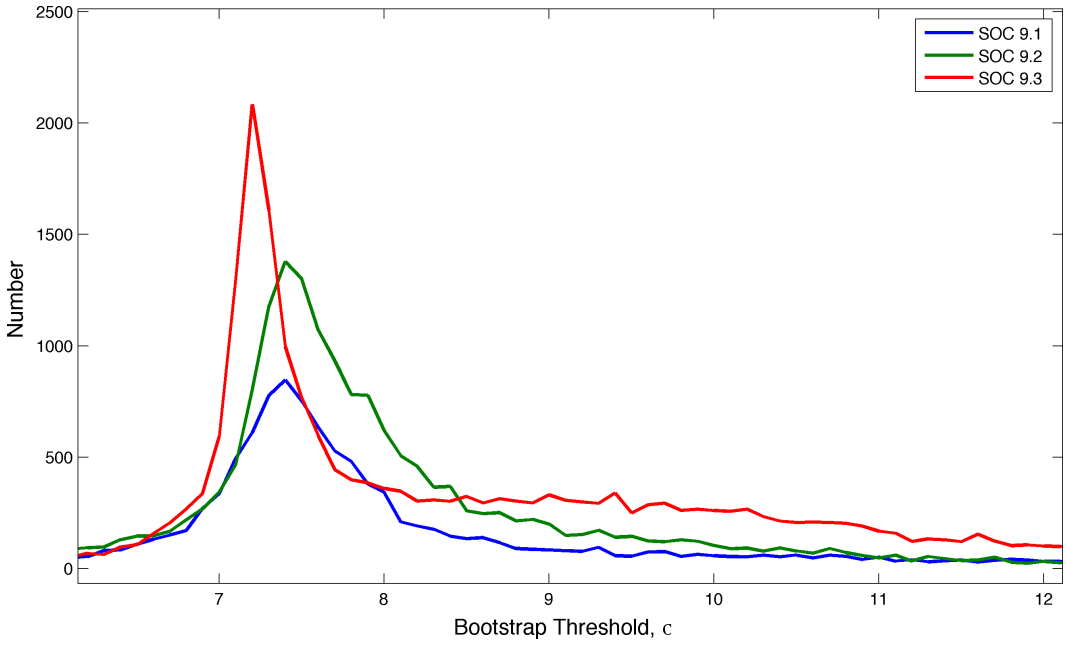

Figure 13. Bootstrap thresholds for KOIs for SOC 9.1 (blue curve), SOC 9.2 (green curve) and SOC 9.3 (red curve).

6. Precision of the Statistical Bootstrap Results

In this section we investigate the precision of the statistical bootstrap test by reviewing the results of a Monte Carlo experiment. The light curve for a typical *Kepler* target observed for all 17 quarters was replaced with a zero-mean, unit variance WGN process and run through TPS to generate the single event statistic time series. This process incorporated all gaps in the original time series in order to present as realistic a result as possible. The single event statistic time series for each of the 14 different pulse durations searched in TPS (between 1.5 and 15 hours) were then subjected to the SOC 9.3 bootstrap analysis. The number of transits, n_{transits}, for each pulse duration was varied between 3 and 2048. A total of 100 random flux time series were generated and subjected to the bootstrap analysis in this manner.

Figure 14 shows the bootstrap false alarm rate at 8σ as a function of the transit duration and the number of transits. For transit pulse durations less than ~3 hours, the false alarm rate is slightly higher than that expected for a fully sampled ZMUV Gaussian process, which would provide a $\log_{10}$(FAR) of -15.2. The FAR then drops gradually to the longest duration, perhaps reflecting the fact that there are more independent statistical tests conducted for shorter transits than for longer duration transits (at a given trial orbital period). The two smallest number of transit cases also exhibit depressed false alarm rates as a function of duration. For eight or more transits, the FAR at 8 σ converges to a rather narrow curve with a width of ~0.25 in log space. For transit durations exceeding 3 hours the dispersion is 2 dex.

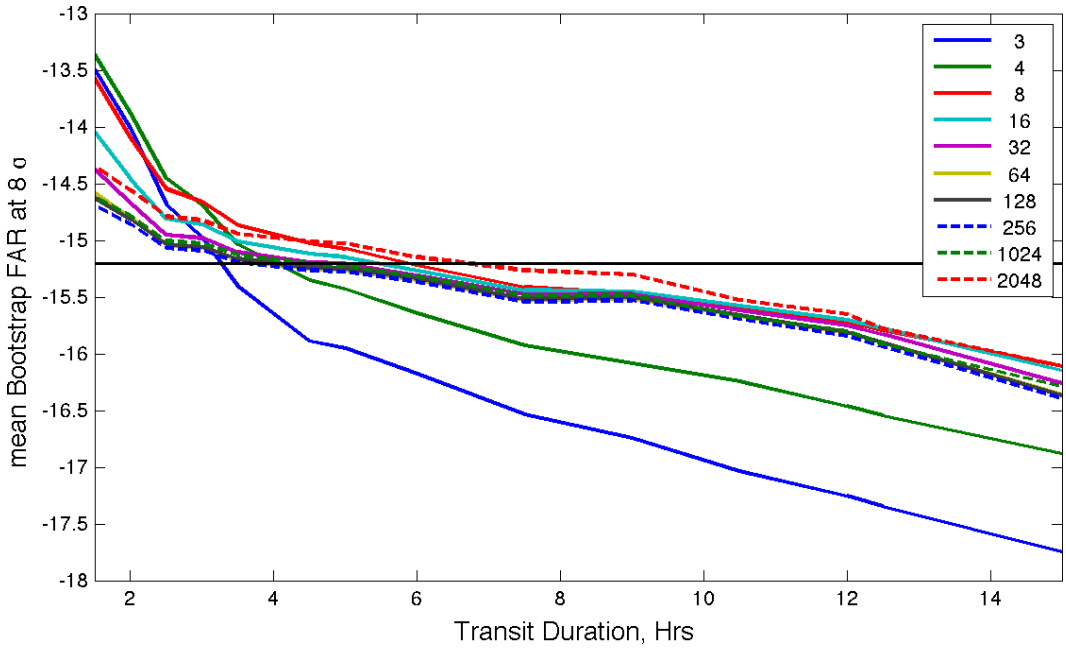

Figure 14. Mean of the log of the false alarm rate at 8 σ of 100 different Monte Carlo tests as a function of transit duration and number of transits. The solid black line indicates the idealized false alarm rate at 8 σ, namely $\log_{10}(6.2\times10^{-16}) = -15.206$.

Figure 15 shows the standard deviation in the $\log_{10}$(FAR) at 8 σ across all 100 Monte Carlo runs as a function of transit duration and number of transits. The curve for $n_{transits}$=3 is between 1.5 and 2.6 over the full range of durations. The curves drop rapidly as a function of $n_{transits}$. For $n_{transits} \geq 8$ the standard deviation of the bootstrap FAR at 8 σ is less than 1 dex. For $n_{transits} \geq 512$, the scatter in the results begins increasing, reflecting accumulation of round off errors due to the spatial resampling that must be used in the bootstrap analysis to prevent spatial aliasing (see Section 3).

Note that the dispersion in the mean bootstrap FAR is comparable to the scatter in Monte Carlo results, indicating that while the mean bootstrap FAR, does, indeed, vary significantly with respect to transit duration and orbital period (number of transits), the bias in an individual bootstrap FAR estimate is approximately the same as the uncertainty, for long orbital periods ($n_{transits} < 8$) and long durations (>6 hr). For these cases, the bootstrap FAR is likely to be biased by 1–2 dex towards smaller false alarm rates.

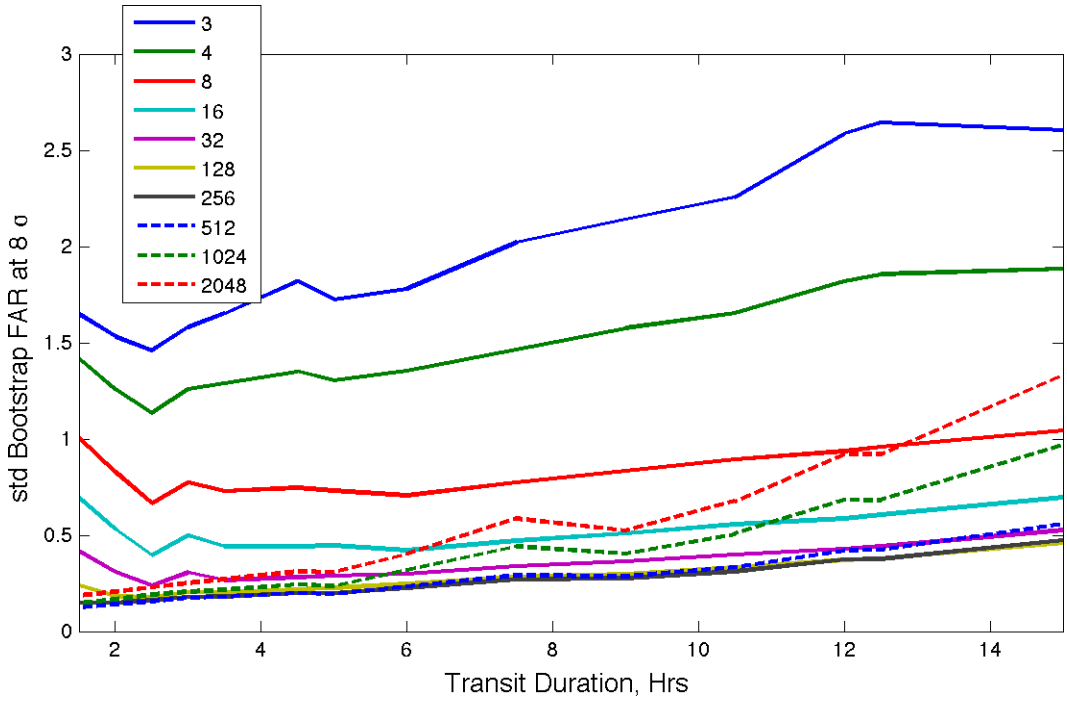

Figure 15. Standard deviation of the log of the false alarm rate at 8 σ of 100 different Monte Carlo tests as a function of transit duration and number of transits. The solid line indicates the ideal threshold at 7.1σ.

Figure 16 shows the mean of the bootstrap threshold as a function of transit duration and $n_{transits}$. The shape of the curves are rather similar to those for the mean $\log_{10}$(FAR) in Figure 14. For durations longer than 3 hours and $n_{transits} \geq 8$ the bootstrap threshold drops linearly from 7.1 σ to 6.6 σ. The curves for $n_{transits} \geq 8$ are relatively tightly confined to within a range of ±0.03 σ. The variation across the curves is ~0.4 σ at any given transit pulse duration, comparable to the scatter of the results across the Monte Carlo trials.

Figure 17 shows the standard deviation of the bootstrap threshold as a function of the transit duration and $n_{transits}$. The shape of the curves is reminiscent of those for the standard deviation of the $\log_{10}$(FAR) in Figure 14. The standard deviation of the threshold is below 0.5 σ for all cases, dropping rapidly from ~0.5 σ for $n_{transits} = 3$ to ~0.2 σ for $n_{transits} = 8$.

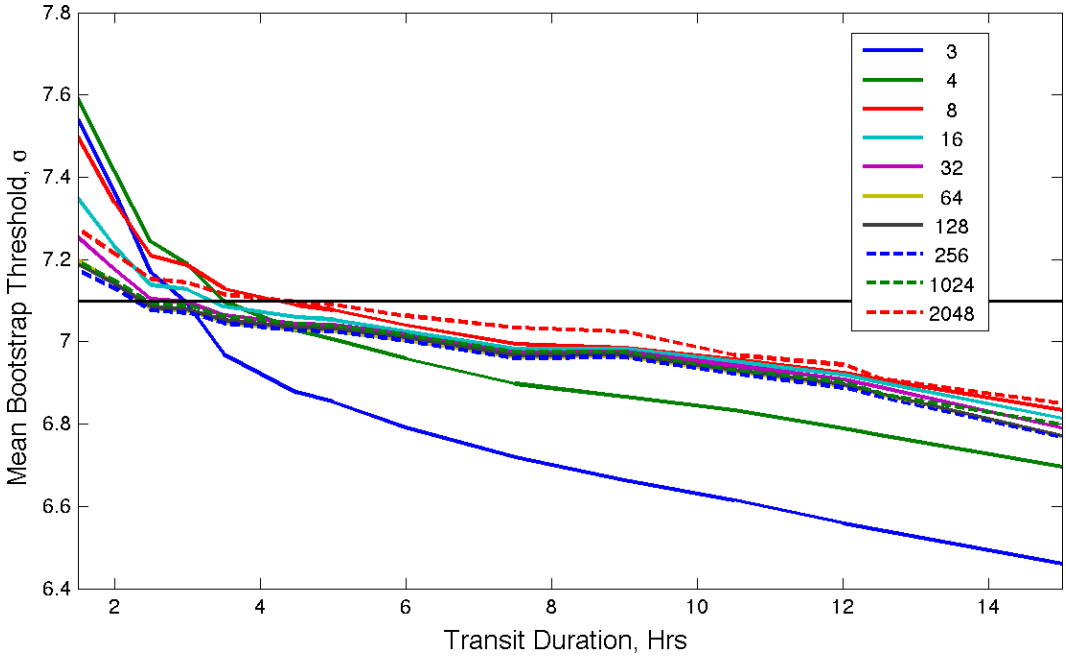

Figure 16. Mean of the bootstrap threshold of 100 different Monte Carlo tests as a function of transit duration and number of transits.

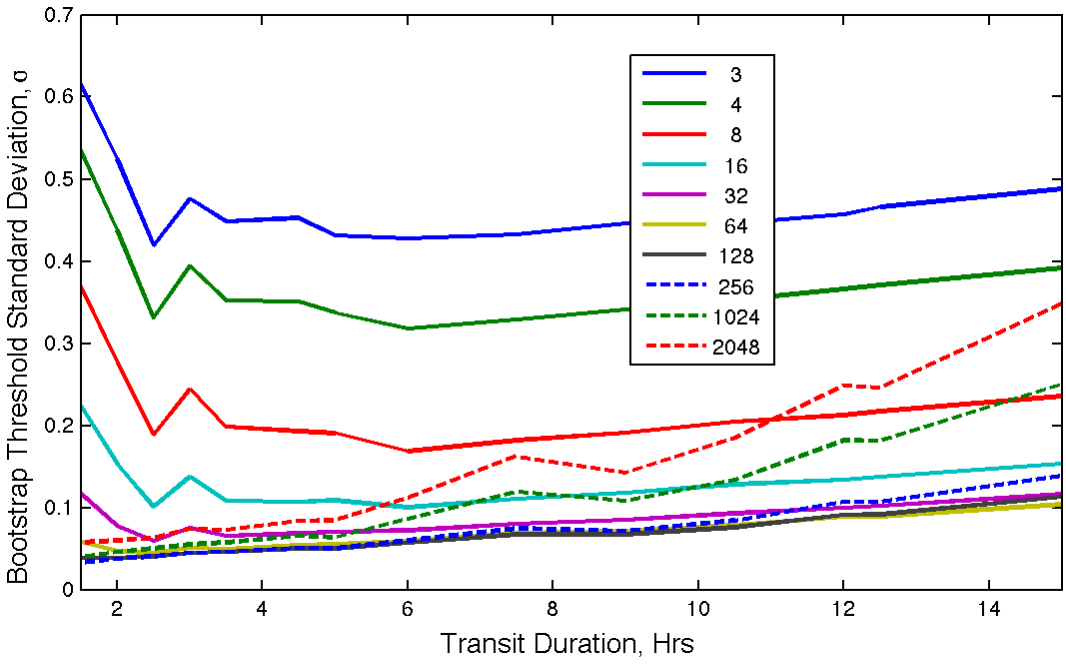

Figure 17. Standard deviation of the bootstrap threshold of 100 different Monte Carlo tests as a function of transit duration and number of transits.

The results are well behaved for transit pulse durations ≥3 hours, and for ≥8 transits. However, the enhanced scatter in the results for small numbers of transits should be kept in mind when interpreting the bootstrap results for TCEs with $n_{transits}$<8.

How much of the bias structure evident in Figures 14 and 16 are due to the bootstrap, and how much is due to the conditioning, filtering and processing within TPS? We conducted a separate Monte Carlo experiment by running bivariate Gaussian multiple event statistics through the bootstrap algorithm directly, thereby bypassing TPS. Figure 18 shows the behavior of the mean bootstrap FAR and threshold at 8σ as a function of the number of transits for this experiment along with those for the original Monte Carlo experiments. The $\log_{10}$(FAR) of the bivariate WGN process varies between -14.5 and -15.4, or +/-1 dex, indicating that the majority of the bias structure dependent on transit duration is due to the filtering and conditioning occurring inside of TPS. We interpret this as being due to the fact that the shorter duration transits have less data "averaged" into each single event statistic, and hence, are noisier than those for longer duration transits, and to the fact that for a finite flux time series, there are more effective independent statistics for shorter duration transits relative to longer duration transits.

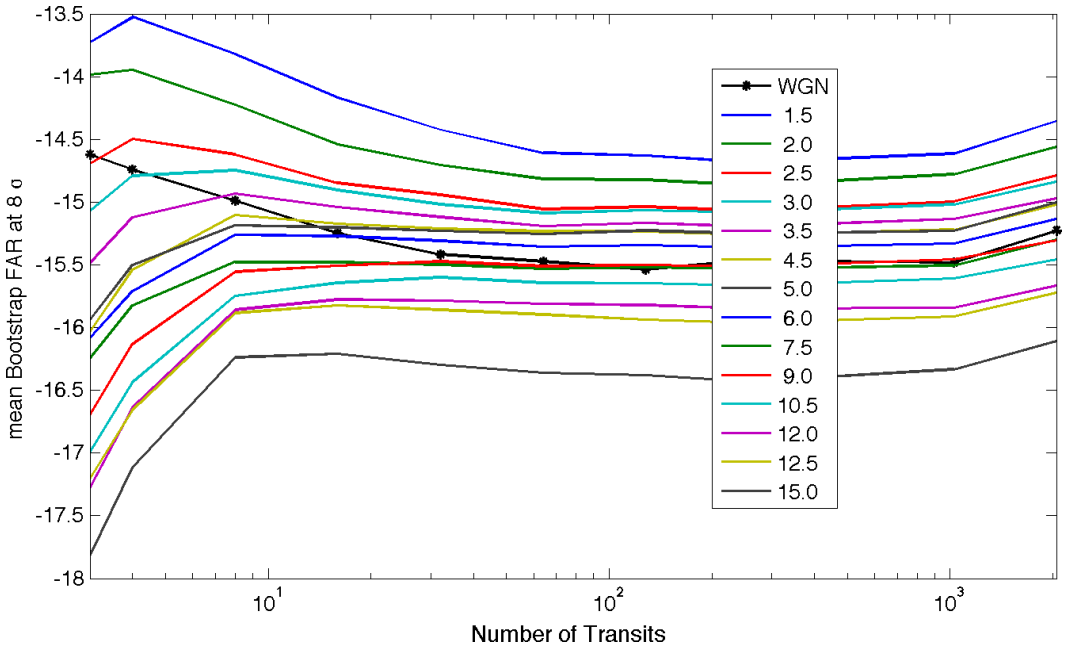

Figure 18. Mean of the bootstrap FAR of 100 different Monte Carlo tests as a function of transit duration and number of transits, along with that for a bivariate WGN process passed directly to the bootstrap algorithm (black curve with stars).

Note that while these Monte Carlo experiments give some idea of the native scatter in the bootstrap analysis results, they do not include all the known instrumental artifacts (e.g. sudden pixel dropouts or rolling band image artifacts) and/or astrophysical red noise.

7. Bootstrap Analysis of a Single TCE

As an illustration of how the statistical bootstrap analysis operates, Figure 19 shows a typical bootstrap result. The TCE used for this plot is on KIC 12158032, which has a transit duration of 2 hours, an orbital period of 0.578 days, and a *MES* of 8.48 σ. If the *MES* of the detection falls above the *MES* corresponding to a $\log_{10}$(FAP) of -13.5, then the boot_fap is interpolated from the CDF of the null *MES* constructed by the bootstrap, otherwise the best-fit Gaussian is used to calculate the boot_fap. In Figure 19 it is marked by the black star and was calculated to be 4.7×10^{-16}. This is the FAP on the solid green curve corresponding to the *MES* of the TCE. The boot_mesthresh for this TCE is ~7.35σ as can be seen by finding the *MES* corresponding to a FAP of $\sim6.24\times10^{-13}$ on the best-fit Gaussian (indicated by the magenta diamond). The boot_mesmean is the mean of the best-fit Gaussian and is -0.64 for this TCE. The boot_messtd is the standard deviation of the best-fit Gaussian and is 1.13 for this TCE. Note that the solid red curve shows the CDF for a ZMUV Gaussian. The Gaussian is fitted robustly in log space using the data from 1×10^{-4} to 1×10^{-13} to avoid the roll-off toward $MES>8\sigma$ due to round off errors.

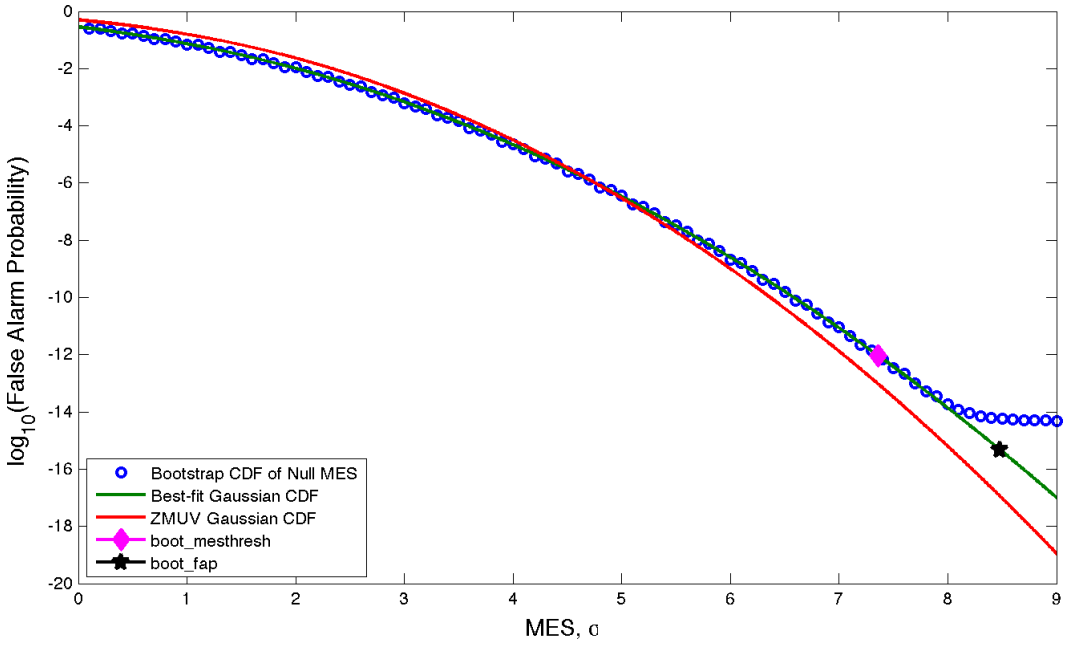

Figure 19. The CDF of the null *MES* constructed by the bootstrap. This TCE is on KIC 12158032 and has a *MES* of 8.48 σ, a duration of 2 hours, and a period of 0.578 days that yielded 2,278 transits in 4 years of data. The false alarm probability for this TCE is ~4.7×10^{-16}, marked by the black star. The best-fit Gaussian had a mean of -0.64 and a standard deviation of 1.13. The magenta diamond marks the threshold needed to achieve the same false alarm rate of a ZMUV Gaussian with a 7.1σ threshold given the distribution of null *MES* constructed by the bootstrap.

8. References

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2016, ApJ, in press, arXiv:1605.05729

Coughlin, J.L., Mullally, F., Thompson, S.E., et al. 2016, *ApJS*, **224**, 12, arXiv:1512.06149

Efron, B. 1979, *The Annals of Statistics*, **7**, 1

Jenkins, J. M. 2002, *ApJ*, **575**, 493

Jenkins, J. M., Caldwell, D. A., & Borucki, W. J. 2002, *ApJ*, **564**, 495

Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010, *Proc SPIE*, **7740**, 77400D

Jenkins, J.M., Twicken, J.D., Batalha, N.M., et al. 2015, *AJ*, **150**, 56

Seader, S., Tenenbaum, P., Jenkins, J. M., & Burke, C. J. 2013, *ApJS*, **206**, 25

Seader, S., Jenkins, J. M., Tenenbaum, P., et al. 2015, *ApJS*, **217**, 18

Smith, J.C., Morris, R. L, Jenkins, J. M. et al. 2016, *PASP*, in press

Tenenbaum, P., Bryson, S. T., Chandrasekaran, H., et al. 2010, *Proc. SPIE*, **7740**, 77400J

Tenenbaum, P., Jenkins, J. M., Seader, S., et al. 2014, *ApJS*, **211**, 6

Twicken, J., Jenkins, J. M., Seader, S. E., et al. 2016, *ApJS*, in press, arXiv:1604.06140

Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010, *Proc. SPIE*, **7740**, 42W