

## THE RADIUS DISTRIBUTION OF PLANETS AROUND COOL STARS

TIMOTHY D. MORTON<sup>1,2</sup>, JONATHAN SWIFT<sup>2</sup>  
*Draft version July 25, 2014*

### ABSTRACT

We calculate an empirical, non-parametric estimate of the shape of the period-marginalized radius distribution of planets with periods less than 150 days using the small yet well-characterized sample of cool ( $T_{\text{eff}} < 4000\text{K}$ ) dwarf stars in the *Kepler* catalog. In particular, we present and validate a new procedure, based on weighted kernel density estimation, to reconstruct the shape of the planet radius function down to radii smaller than the completeness limit of the survey at the longest periods. Under the assumption that the period distribution of planets does not change dramatically with planet radius, we show that the occurrence of planets around these stars continues to increase to below  $1 R_{\oplus}$ , and that there is no strong evidence for a turnover in the planet radius function. In fact, we demonstrate using many iterations of simulated data that a spurious turnover may be inferred from data even when the true distribution continues to rise toward smaller radii. Finally, the sharp rise in the radius distribution below  $\sim 3 R_{\oplus}$  implies that a large number of planets await discovery around cool dwarfs as the sensitivities of ground-based transit surveys increase.

### 1. INTRODUCTION

The discovery of the first exoplanets (Wolszczan & Frail 1992; Mayor & Queloz 1995; Marcy & Butler 1996) has sparked tremendous growth in research and interest in the formation and evolution of planetary systems beyond the Solar System. Most of the first few dozen extrasolar planets found had masses greater than Saturn and semimajor axes less than 0.5 AU. But not unlike many areas of astronomy, the first discoveries are not representative samples; rather, close-in giant planets are relatively rare (Wright et al. 2012; Howard et al. 2010) in comparison to the new populations of exoplanets now being revealed by the *Kepler* Mission (Borucki et al. 2011; Batalha et al. 2012; Burke 2013). The most common kinds of planets within *Kepler*'s discovery space of  $R_p \gtrsim 0.5R_{\oplus}$ , and  $P \lesssim 100\text{d}$  appear to be somewhat larger than Earth but smaller than Neptune,  $1 < R_p < 4R_{\oplus}$  (Howard et al. 2012; Fressin et al. 2013; Dressing & Charbonneau 2013).

Much of our understanding of planet formation is anchored in decades of research into our own Solar System. But now the burgeoning exoplanet population provides us with a new context revealing important insights into planet formation throughout the Galaxy. For example, the large amount of planetary mass seen close to host stars is evidence that protoplanetary disks may have much higher surface densities than previously thought (Hansen & Murray 2012; Chiang & Laughlin 2012) or that the observed planets migrated from regions further from their host star where more mass was readily available for assembly (Swift et al. 2013; Raymond & Cossou 2014).

The architectures of *Kepler* planetary systems also offer a wealth of information regarding their formation and evolution. The masses of a subset of planets in *Kepler*'s

multi systems have been measured from the effects of mutual gravitational interactions (e.g. Cochran et al. 2011, Lithwick et al. 2012, Nesvorny et al. 2013) providing insight into the composition and atmospheric evolution of these planets (Wu & Lithwick 2013; Owen & Wu 2013; Lopez et al. 2012). While more recently, individual planet masses are being measured with precise radial velocity observations (Marcy et al. 2014), advancing our understanding of the compositional makeup of the *Kepler* planet sample (e.g., Weiss & Marcy 2014). The low inferred mutual inclination of multi-transit systems ( $\sim 1^{\circ}$ – $3^{\circ}$ ; Fabrycky et al. 2012; Fang & Margot 2012) together with the relative number of single versus multi-transit systems provides constraints on the number of planets in a given system within *Kepler*'s discovery window, else it may be the first indication of a separate, high-inclination population of single transit systems (Hansen & Murray 2013; Fang & Margot 2012).

In this article we focus on yet another important clue regarding the formation of the compact systems revealed by *Kepler*: the distribution of planetary radii. Constructing the underlying distribution of planet radii from the results of a transit survey is a subtle task. Early in the history of transit surveys, Gaudi (2005) and Gaudi et al. (2005) identified many of the pitfalls inherent to this endeavor—mainly the strong period and radius biases. While the characteristics of the *Kepler* survey make it less susceptible to these issues, these biases still persist and need to be accounted for.

The radius distribution of planets as derived from the *Kepler* survey has been the subject of a number of studies in the literature. The initial estimates of the planet radius distribution by Howard et al. (2012) showed a dramatic increase in the number of planets at ever smaller size. Citing incompleteness, however, they did not follow this trend in their analysis to planet radii smaller than  $2 R_{\oplus}$ . In an independent study by Youdin (2011), a parametric estimation of the planetary distribution function revealed a deficit of large planets in short period orbits that would support a formation scenario involving core accretion followed by inward migration; Dong & Zhu

tdm@astro.princeton.edu

<sup>1</sup>Department of Astrophysical Sciences, 4 Ivy Lane, Peyton Hall, Princeton University, Princeton, NJ 08544

<sup>2</sup>Department of Astrophysics, California Institute of Technology, MC 249-17, Pasadena, CA 91125

(2012) report a similar finding.

More recent estimates of the planet radius distribution show a preferential size scale in the *Kepler* sample indicated by a flattening and possible turnover in the log-binned histogram of detected planet candidates somewhere around  $2 R_{\oplus}$  (Fressin et al. 2013; Dressing & Charbonneau 2013; Petigura & Marcy 2013; Petigura et al. 2013). If true, this would be an important clue toward understanding the key mechanisms that shape the observed population of compact planetary systems that pervade the Galaxy. However, these analyses are constrained by the limitations of coarse histograms; no analysis to date has yet characterized the shape of the exoplanet radius distribution in enough detail to allow meaningful comparison to planet formation and evolution theories. Additionally, all of the above except for Dressing & Charbonneau (2013) rely on the *Kepler* Input Catalog for the stellar properties of the target star population, the known unreliability of which has been shown to bias the results of statistical analyses (Gaidos & Mann 2013).

The goal of the present study is distinct from previous work in several ways. First, we focus exclusively on the smallest stars in the *Kepler* target sample as a well characterized subsample, due to both the photometric re-calibration of stellar properties presented by Dressing & Charbonneau (2013) and because the host stars of many of the *Kepler* Objects of Interest (KOIs) in this sample have been investigated spectroscopically (Muirhead et al., in prep; Muirhead et al. 2012b). Second, we aim to reconstruct as faithfully as possible the detailed shape of the radius function, avoiding both the limitations of histogram binning and the assumption that the distribution follows a power law—a non-parametric, non-histogram approach to this problem has not yet been attempted. And finally, in order to investigate any potential flattening or turnover of the distribution, we extend planet occurrence analysis to radii smaller than has been attempted before, introducing a new technique that allows proper marginalization over period even for radii for which the *Kepler* survey is beginning to be incomplete.

In §2 we walk through the steps required to extract this non-parametric empirical estimate of the true period-marginalized planet radius function given a population of detections from a well-characterized transit survey. In §3 we apply these methods to the Cool KOIs to derive the radius distribution for small planets around small stars down to below  $1 R_{\oplus}$ . We summarize our results in §4 and explore the assumptions that go into this calculation in §5. Concluding remarks follow in §6.

## 2. FORMALISM

We define the planet radius distribution function  $\phi_r^{P_{\max}}(r)$  such that

$$\int_{r_{\min}}^{r_{\max}} \phi_r^{P_{\max}}(r) dr = \text{NPPS}, \quad P < P_{\max}; \quad (1)$$

that is, a density function with an overall normalization giving the average number of planets per star (NPPS) for planets with period less than  $P_{\max}$  days, for planet radii  $r$  between  $r_{\min}$  and  $r_{\max}$ . The problem of calculating planet occurrence rates from *Kepler* has been quite an industry over the last few years (Youdin 2011; Howard et al. 2012; Dong & Zhu 2012; Swift et al. 2013; Fressin

et al. 2013; Petigura & Marcy 2013; Dressing & Charbonneau 2013). However, there has been little quantitative discussion of deriving the detailed shape of the radius function beyond drawing histograms. In the following subsections, we review and refine the general principles of an occurrence calculation and then describe how to follow these principles to construct a non-parametric empirical radius function that obeys the above desired properties.

### 2.1. Occurrence Calculations

In a perfectly idealized survey that is both 100% reliable and 100% complete, the occurrence rate of planets (in a survey, or in a specified bin) is simply

$$\text{NPPS} = \frac{N_p}{N_{\star}}, \quad (2)$$

where  $N_p$  is the number of detected planets and  $N_{\star}$  is the number of stars surveyed. In practice, however, this must be corrected for both incompleteness and unreliability as follows:

$$\text{NPPS} = \frac{1}{N_{\star}} \sum_{i=1}^{N_p} w_i. \quad (3)$$

Here the sum is over all detections and  $w_i$  is a weighting factor applied individually to account for the various necessary corrections. Generally, these weights can be thought of as

$$w_i = \frac{(1 - \text{FPP}_i)}{\eta_i}, \quad (4)$$

where  $\text{FPP}_i$  is the probability that signal  $i$  is a false positive and  $\eta_i$  is an individualized efficiency factor for the detection of planet  $i$ . In this work, we calculate the  $\text{FPP}_i$  according to the procedure in Morton (2012); see §3.4 for details.

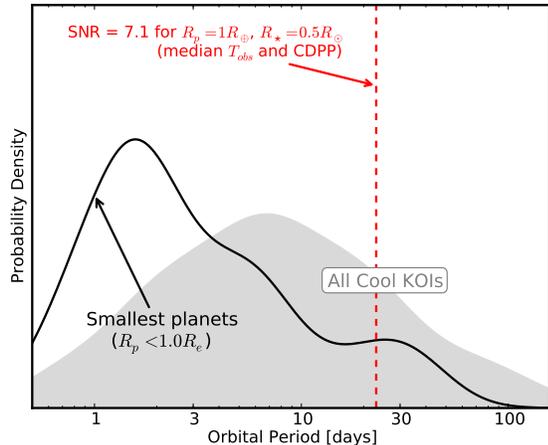
We thus focus discussion here on the detection efficiency  $\eta_i$ , which is defined by the following thought experiment: *If a very large number of planets identical to planet  $i$  were distributed randomly around all the stars in the survey, only a fraction  $\eta_i$  could have been detected.* Conceptually, this can be factored (following Youdin 2011):

$$\eta_i = \eta_{\text{tr},i} \cdot \eta_{\text{disc},i}, \quad (5)$$

where  $\eta_{\text{tr}}$  is the geometric transit probability, and  $\eta_{\text{disc}}$  is the “discovery efficiency”: the fraction of planets in this thought experiment *with transiting orbital geometries* that could have been detected by the survey. Previous *Kepler* occurrence rate calculations (Howard et al. 2012; Swift et al. 2013; Dressing & Charbonneau 2013), have defined the discovery efficiency as

$$\eta_{\text{disc},i} = \frac{N_{\star,i}}{N_{\star}} \quad (6)$$

where  $N_{\star,i}$  is the number of target stars around which planet  $i$  could have been detected. This number is determined by counting the stars around which a transit of planet  $i$  (at period  $P_i$ ) would cause a photometric signal with signal-to-noise ratio (SNR) above some detection threshold (10 for Howard et al. (2012), 7.1 for Dressing & Charbonneau (2013) and Swift et al. (2013)).



**Figure 1.** Evidence supporting the hypothesis that small planets are incomplete in the Cool KOI sample. The solid black line is the observed period distribution of planets smaller than  $1 R_{\oplus}$ ; the grey shaded area is the observed period distribution of all the Cool KOIs. (Neither distribution is corrected for transit probability.) The vertical dashed red line indicates the period at which a  $1 R_{\oplus}$  planet around a  $0.5 R_{\odot}$  star (typical of the Cool KOI sample) would have SNR of 7.1, the nominal detection threshold for KOI identification. The relative lack of observed small planets at periods longer than 10 days is thus very plausibly due to incompleteness.

Though not spelled out explicitly, this way of calculating  $\eta_{\text{disc},i}$  essentially boils down to two general ingredients: a hypothetical “ensemble-of-alternative-scenarios” SNR probability distribution  $\phi_{\text{SNR},i}$  for each planet in the survey, and the discovery efficiency as a function of SNR  $\eta_{\text{SNR}}$  for the transit detection pipeline. Given these two ingredients, Howard et al. (2012), Dressing & Charbonneau (2013), and Swift et al. (2013) all implicitly calculate the following:

$$\eta_{\text{disc},i} = \int_0^{\infty} \eta_{\text{SNR}}(s) \cdot \phi_{\text{SNR},i}(s) ds, \quad (7)$$

where  $\phi_{\text{SNR},i}$  is a normalized probability distribution of SNR ( $s$ ) for that planet determined by varying the stellar radius and noise properties according to each star in the survey, and  $\eta_{\text{SNR}}$  is assumed to be a step function at some detection threshold.

It is important to note, however, that a step function does not accurately characterize the true detection efficiency of the *Kepler* pipeline as a function of SNR. Fressin et al. (2013) found that the observed SNR distribution of the Batalha et al. (2012) catalog implied that the true behavior was more like an “SNR ramp,” where the discovery efficiency was zero at  $\text{SNR} = 6$  and unity at  $\text{SNR} = 16$ . Internal tests of the *Kepler* pipeline indicate that the true shape of the  $\eta_{\text{SNR}}$  function is similar to this (P. Tennenbaum, priv. comm., based on poster at the 2nd *Kepler* Science Conference). An accurate estimate of this function is crucial to obtaining correct results in an occurrence rate calculation; we present the functional form we use in §3.

We also note that the traditional method of computing  $\phi_{\text{SNR},i}$ —simulating a planet of radius  $r_i$  transiting all of the survey stars at fixed period  $P_i$ —is designed to estimate the *joint period-radius distribution* of planets. This joint distribution is then summed up over period to

find the period-marginalized radius distribution—this is how planet occurrence as a function of radius is determined in Howard et al. (2012), Dressing & Charbonneau (2013), Petigura & Marcy (2013), and Petigura et al. (2013). However, as Figure 1 demonstrates, the smallest planets in the Cool KOI survey are likely only complete out to periods of a few tens of days, whereas many larger planets are detected on larger orbits. Thus, determining the period-marginalized radius function of planets using this traditional approach requires either restricting the analysis to radii above which the survey is substantially complete out to  $P = P_{\text{max}}$ , or only using a small fraction of the survey detections (out to periods of only a few tens of days) in order to probe below  $\sim 1 R_{\oplus}$ .

We introduce here an alternative method of computing  $\phi_{\text{SNR},i}$  that allows for better reconstruction of the planet radius function at small radii without having to restrict analysis to a fraction of the available data. Instead of simulating the ensemble of alternative transit scenarios for planet  $i$  all at the single fixed period  $P_i$ , we suggest that these alternative scenarios could be assigned a distribution of periods according to a reasonable estimate of the true planet period distribution. This effectively amounts to a strategy of “pre-marginalization” that requires an assumption of the period distribution of planets but allows for every planet detection to be treated on an equal basis.

Making this adjustment to how  $\eta_{\text{disc},i}$  is determined—by distributing hypothetical planets around all stars *at all periods* to calculate  $\phi_{\text{SNR},i}$ —also necessitates rethinking how transit probability is corrected for. That is, construction of  $\phi_{\text{SNR},i}$  must acknowledge that transit probability (as well as SNR) depends on planet period and host star radius. In other words, any process of building up a hypothetical SNR distribution from many instances of simulated transits must ensure that each instance is properly weighted by its actual likelihood of detection. The most straightforward way of ensuring this is to simulate an isotropic cloud of planets around each target star and include only the transiting configurations (e.g., impact parameter  $b \leq 1$ ) in the SNR distribution. The transit probability factor  $\eta_{\text{tr},i}$  then becomes the fraction of all these simulated planets that transit—this will typically *not* be the same as the individual geometric transit probability of planet  $i$ , since planet  $i$  has a single period  $P_i$  while the simulated population has a distribution of periods. In other words, calculating the radius function pre-marginalized over period in this way demands that only a single survey- and period-averaged transit probability be used when calculating the completeness correction for each planet—a function of only the assumed period distribution and the distribution of stellar radii of the target sample.

## 2.2. Estimating the Radius Distribution Function

In all the *Kepler* planet occurrence calculations to date, the shape of the radius function has been explored only very coarsely, by calculating the occurrence rate in several different radius bins and either fitting a power law or qualitatively commenting on the shape. Howard et al. (2012) found a good fit to an  $R^{-2}$  power law down to  $2 R_{\oplus}$ , and declined to comment for smaller planets. On the other hand, Fressin et al. (2013) and Petigura & Marcy (2013) note that the occurrence rate of planets

increases towards smaller radius but then appears to flatten out below about  $2.8 R_{\oplus}$ . Dressing & Charbonneau (2013) claim that the occurrence rate begins to decrease for planets smaller than  $1-1.4 R_{\oplus}$ , and Petigura et al. (2013) find that the planet occurrence rate decreases for planets smaller than  $2 R_{\oplus}$ .

Investigating the shape of the radius distribution in more detail requires a non-parametric approach, and also should avoid binning. Here we introduce the concept of using a weighted kernel density estimator in order to accomplish this.

### 2.2.1. Weighted Kernel Density Estimation

A standard kernel density estimator (KDE) attempts to estimate the true underlying probability distribution of a sample of data points using a function of the following form:

$$\hat{\phi}(x) = \frac{1}{N} \sum_{i=1}^N k(x - x_i; \sigma_i), \quad (8)$$

where  $N$  is the number of data points and  $k(x)$  is a zero-mean, normalized kernel function of arbitrary shape (commonly a Gaussian), with some bandwidth  $\sigma_i$ , that most generally can be different for each data point. This creates a smooth distribution out of a discrete data set, with the degree of smoothness controlled by the width parameter. The choice of width has tradeoffs in both directions: if the kernels are too narrow the estimator will be bumpy, but if they are too wide they can wash out real structure in the distribution. Often the width is selected to be the same for all points based on the number of data points, or sometimes a variable-width kernel is used, e.g. the distance to the  $n$ th nearest neighbor. The  $1/N$  normalization factor assures that the integral of this density estimator over the whole parameter space is unity.

In order to use the KDE concept to properly reconstruct the radius function of planets detected in a transit survey, each data point has to be weighted appropriately, leading to a weighted KDE, or wKDE:

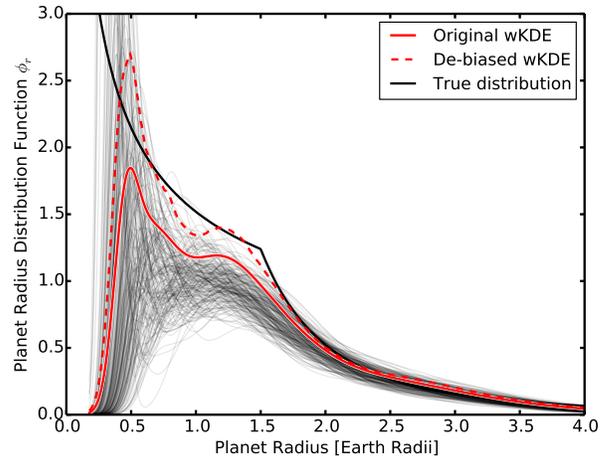
$$\hat{\phi}_r^{P_{\max}}(r) = \frac{1}{N_{\star}} \sum_{i=1}^{N_p} w_i \cdot k(r - r_i; \sigma_i), \quad (9)$$

where  $w_i = 1/\eta_i$  are the appropriately calculated individual weight factors that renormalize the kernels to correct for missing planets, as discussed in §2.1. The weights ensure that the shape of the radius function responds appropriately to the individual corrections, and the  $N_{\star}$  overall normalization ensures that the integral over all radii will return the NPPS, as desired in Equation 1. We note that because this function does not normalize to unity, it is not strictly a density function, but a “rate function,” representing number of planets per unit radius rather than probability density.

Wang & Wang (2007) explore in detail several techniques for selecting the optimal kernel bandwidths, presenting in particular two methods of the form

$$\sigma_{\text{opt}} = 0.9Cn^{-1/5} \quad (10)$$

where  $C$  is a constant based on the sample mean, sample variance, or interquartile range of the data, and  $n$  is the



**Figure 2.** An illustration of the de-biasing procedure for the wKDE estimator discussed in this work. Mock transit survey data is simulated, with the size of the planets following the distribution illustrated by the thick black line. From this data, a wKDE of the radius distribution is derived using the methods described in §2.2.1 (solid red line). Then, 200 new transit-survey datasets are generated with the planet radii drawn from the estimated red-line distribution, and corresponding new wKDEs are generated according to the same procedure (thin black lines). The thin black lines tend to fall systematically below the solid red line; this is an illustration of the bias of this estimator. This difference is added back to the red line to produce the de-biased estimate of the distribution (dashed red line), which matches the true distribution more closely. The spread in the bootstrap estimators around their median illustrates the estimator’s variance. This technique is known as the “smoothed bootstrap” (Narsky & Porter 2013).

number of data points. However, we show in §3 that for our particular case, using the actual observational uncertainties in planet radius to control the width of the individual smoothing kernels  $\sigma_i$  results in a smooth distribution, so we do not apply this technique. We do note, however, that if the planet radii were known more precisely, then a method such as this would be necessary in order to quantify an optimum smoothing width to avoid high-frequency features in the estimator.

### 2.2.2. De-biasing the wKDE and calculating variance

The wKDE  $\hat{\phi}_r$  described in Equation 9 is an estimator of the true underlying radius distribution  $\phi_r$ , and like any estimator, it has both bias and variance associated with it. These quantities must be determined in order to best understand what the data can tell us about the true distribution. Narsky & Porter (2013) explain how to calculate bias and variance for a standard KDE using a resampling technique called the “smoothed bootstrap”; we adopt this method for our purpose and describe it here.

While a traditional bootstrap technique involves resampling the observed dataset with replacement to create new datasets, a “smoothed bootstrap” involves generating new datasets *according to the estimated distribution*. The density estimator is recalculated for each of these simulated datasets using the same procedure that generated the original estimator. The median offset of these resampled estimators relative to the the original estimator (which can be directly observed) will then reflect the bias of the original estimator relative to the true underlying distribution, and the original estimator can thus

be corrected accordingly. Similarly, the variance of the estimator can be determined by the scatter of these bootstrapped estimators.

Applying this method to a wKDE in the context of a transit survey is less straightforward than simple resampling according to the derived  $\hat{\phi}_r$ , but we borrow the same principle. Each smoothed bootstrap resampling dataset is generated by simulating a whole new survey: drawing a set of planets according to the estimated distribution, assigning them isotropically distributed orbits around host stars drawn randomly from the target sample, calculating the SNR of each of the resulting transits, and then determining which of these planets would be detectable (using the appropriate  $\eta_{\text{SNR}}$  function). Each of these datasets is then converted into a new wKDE, and the bias may thus be assessed.

Figure 2 illustrates this procedure applied in a toy scenario. The true underlying planet radius distribution—a broken power law—is illustrated with the thick solid black line. The initial wKDE estimate of the radius distribution derived from one realization of “detected” planets is shown as a solid red line. Thin black lines illustrate 200 wKDEs resulting from resampled datasets generated according to the originally calculated wKDE distribution (red line). It is clear that for below  $1.5 R_{\oplus}$  or so, the bootstrap resamplings underestimate the red line; this mimics the way the red line is biased with respect to the true distribution. So to correct for this, the difference between the solid red line and the median of the bootstrapped distributions is added to the the initial wKDE estimate at each value of radius to obtain the de-biased wKDE (dashed red line), which matches the true distribution more closely. The error region around this de-biased estimate is then taken to be the same as the spread in the bootstrapped wKDEs about their median.

In addition to illustrating how the bias and variance of these wKDEs may be calculated, this example also touches on two other important points. First, this procedure demonstrates that naive summing of weights to calculate NPPS (Equation 2) is in fact most generally a *biased estimator* for the true NPPS—a non-intuitive but significant result. And second, the toy-model true radius distribution used here is a power law that continues to rise all the way down to  $0.3 R_{\oplus}$ . However, because of the detection sensitivity of the survey, only very rarely are any planets smaller than  $0.5 R_{\oplus}$  detected. Thus, there is generically a turnover in the estimated planet radius distribution—the location of which depends on which planets happen to be observed. In fact, the estimated distribution sometimes even appears to turn over around  $\sim 1 R_{\oplus}$ —despite the fact that the true distribution continues to rise. Recognizing this is crucial to properly interpreting the radius function derived from *Kepler* data; we return to discussion of this point in §5.

### 3. CALCULATING THE COOL KOI RADIUS FUNCTION

One of the biggest concerns to date about interpreting *Kepler* data is uncertainty about stellar parameters. This applies both because the properties of the transit host stars are unknown (derived planet radius depends directly on the radius of the host star) and because the properties of the stars in the survey parent sample are unknown (i.e. is *Kepler* actually surveying dwarf stars or is the parent sample significantly contaminated by giants

or subgiants? (Mann et al. 2012)).

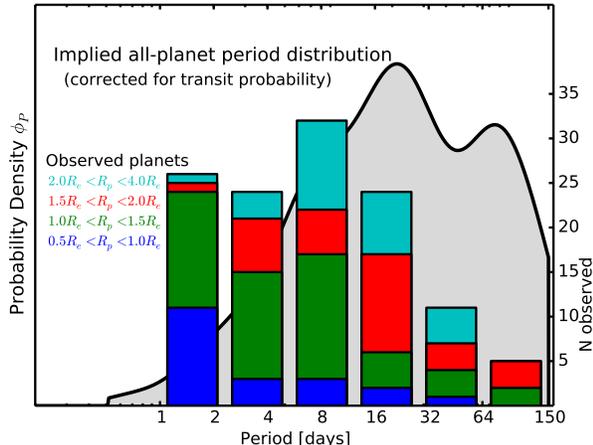
Focusing on *Kepler* candidates around relatively low-mass stars alleviates these concerns. Many of these stars have spectroscopically measured stellar properties (Muirhead et al. 2012a; Mann et al. 2012), and in addition, the properties of the parent sample of target stars has been carefully characterized photometrically by Dressing & Charbonneau (2013). Such an investigation thus is narrower than attempting to use the whole *Kepler* sample, but the assurance of a good understanding of the stellar parameters of both the host stars and the general survey sample more than compensates for this loss of generality. In addition, focusing on these “Cool KOIs” enables detailed study of the radius distribution of Earth-sized and smaller planets.

To construct the planet radius function, we thus select the 130 *Kepler* Objects of Interest (KOIs) with periods  $< 150\text{d}$  identified in the Q1-Q12 KOI catalog posted at the NASA Exoplanet Archive that are hosted by stars with  $T_{\text{eff}} < 4000\text{K}$  as characterized by Dressing & Charbonneau (2013). To this stellar sample we add KOI-961/Kepler-42, which was left out of the Dressing & Charbonneau (2013) sample because its broad-band colors are consistent with classification as either a giant or a dwarf, even though it has been spectroscopically confirmed to be a  $\sim 0.15 M_{\odot}$  dwarf (Muirhead et al. 2012b). (We do note, however, that KOI-961.02 is not in this Q1-Q12 KOI catalog because the minimum period in the Q1-Q12 *Kepler* pipeline search was 0.5 days.) For stellar parameters we use the results presented in Dressing & Charbonneau (2013), except for those KOI host stars that have been spectroscopically characterized according to the observations and procedures described in Muirhead et al. (2012b), for which we use the spectroscopic parameters (Muirhead et al., in prep).

In the following subsections, we describe the steps necessary to calculate  $\hat{\phi}_r^{150}$ , the estimate of the radius function for planets on orbits  $< 150\text{d}$ , from this KOI sample. As described in §2, the crucial step toward properly estimating the radius function is calculating the weight factor  $w_i = 1/\eta_i$  for each detection, which includes a transit probability factor and a completeness factor  $\eta_{\text{disc},i}$  (Equation 7). Key to calculating  $\eta_{\text{disc},i}$  is determining the SNR distribution of a hypothetical population of clones of planet  $i$  around all the target stars, or  $\phi_{\text{SNR},i}$ . To accomplish this, we use the “pre-marginalization” strategy as explained in §2.1, which in turn requires an assumption of the intrinsic period distribution of planets  $\phi_P$ .

#### 3.1. Period distribution

In order to estimate the shape of the true period distribution of planets of all sizes, we make the simplifying assumption that the period distribution of planets is independent of their radii (see §5.1 for discussion regarding this assumption). We thus construct the distribution of  $\log P$  from all the planet candidates in the sample, using a wKDE as described in §2.2. For the weights we use only the inverse transit probabilities, and enforce that the whole distribution is normalized to unity, creating the probability density function for  $\log P$ . For the widths we use  $\sigma = 0.15$  (in  $\log P$ ), to create a smooth distribution. This is the period distribution function  $\phi_P$  that we



**Figure 3.** The period distribution of planets around *Kepler*’s M dwarfs. The grey shaded region is the implied period distribution of all planets combined, correcting for the effects of transit probability. The bar charts show the observed numbers of planets of different sizes in each period bin. Note the declining fraction of small planets as a function of period—this is most likely an effect of declining detection efficiency for smaller planets on longer-period orbits, and this must be properly accounted for when constructing the period-marginalized planet radius function down to small radii. The radius function calculation in this paper assumes that all planets are distributed according to the shaded distribution, regardless of planet radius. See §5.1 for a discussion of this assumption.

use in the following subsection, shown as the grey shaded region in Figure 3.

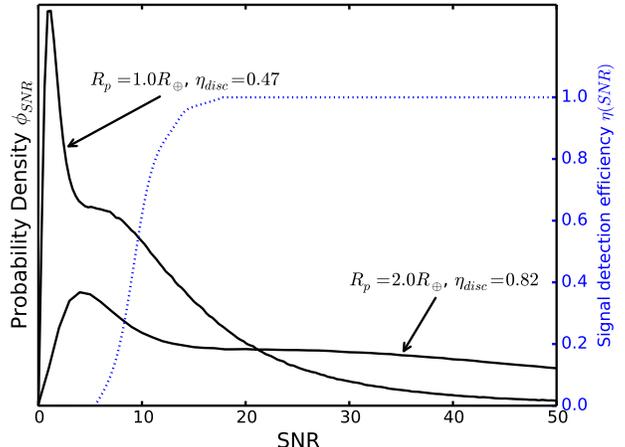
### 3.2. SNR distribution

The SNR of a transit signal is usually defined as follows:

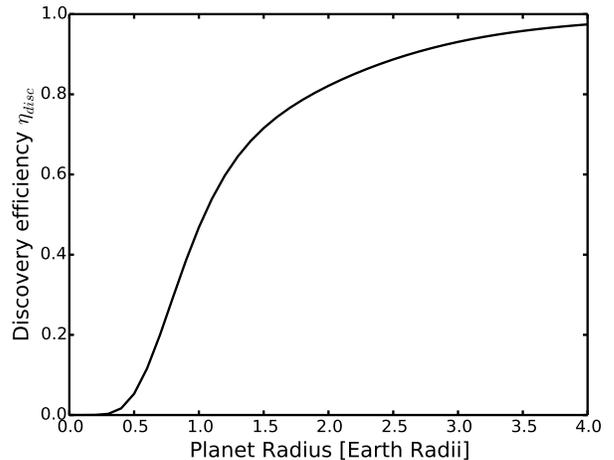
$$\text{SNR} = \frac{\delta}{\sigma} \sqrt{N_{\text{tr}} \cdot N_{\text{pts}}}, \quad (11)$$

where  $\delta$  is the transit depth,  $\sigma$  is the one-point photometric uncertainty,  $N_{\text{tr}}$  is the number of transits observed, and  $N_{\text{pts}}$  is the number of photometric points per transit.

In order to construct  $\phi_{\text{SNR},i}$ , the distribution of SNRs for every conceivable alternative scenario in which planet of radius  $r_i$  might have existed in this survey (around every other target star at any other potential period, under our pre-marginalization strategy; see §2.1), we use a Monte Carlo simulation. For each target star, 50,000 planets of radius  $r_i$  are generated with isotropic orbital inclinations and orbital periods according to the period distribution of §3.1. We then calculate the transit signal SNR for each planet in this simulation that has a transiting impact parameter, with  $\delta$  being the depth of a Mandel & Agol (2002) transit model around that star averaged over the transit duration (using limb darkening parameters for each target star from Claret et al. (2012)),  $\sigma$  being the published median three-hour combined differential photometric precision (CDPP) for that star,  $N_{\text{pts}}$  being the the total transit duration  $T_{\text{dur}}/(3 \text{ hr})$ , and  $N_{\text{tr}}$  being the orbital period of the simulated planet divided by the total amount of time that star was observed by *Kepler* (number of quarters up until Q12,  $\times 90$  days, excepting Q1, which is 33 days). We note that while this formulation ignores details of the noise properties on the exact timescale of transit and the exact observing window function, this prescription for cal-



**Figure 4.** Two examples of the SNR distributions resulting from simulating transits of a planet of a certain size around every *Kepler* target star considered in this study, randomly assigning periods and impact parameters. The properties of this distribution depend on the radii and photometric noise levels of the target stars, and the integral of the product of the pipeline detection efficiency function (blue dotted curve, from P. Tennenbaum, priv. comm.) with this distribution gives the “discovery fraction”  $\eta_{\text{disc}}$  (Equation 7). A  $1 R_{\oplus}$  planet is detectable in only about half of potential transiting configurations, whereas a  $2 R_{\oplus}$  planet is detectable in about 4/5 of potential configurations.



**Figure 5.** The discovery efficiency of the *Kepler* pipeline as a function of planet radius for the subset of *Kepler* target stars considered in this work. At each value of planet radius,  $\eta_{\text{disc}}$  is calculated according to Equation 7, using the SNR distributions as calculated in §3.2 and illustrated in Figure 4.

culating SNR is exactly the use case of the CDPP values as recommended by the *Kepler* team (Christiansen et al. 2012), especially since most of the targets in this survey are faint and thus white-noise dominated on the timescale of transits. 2.6% of the planets simulated in this exercise—50,000 for each target star—have transiting geometries; this is the survey- and period-averaged transit probability discussed in §2.1.

Rather than repeat this exact procedure for every planet, it is sufficient to generate the distribution once for a fiducial planet radius  $r_0$  (e.g.,  $1 R_{\oplus}$ ), and then create the distribution appropriate for any other planet with radius  $r$  by multiplying all the SNRs in the fiducial

distribution by  $(r/r_0)^2$ . Figure 4 illustrates these distributions for  $1 R_\oplus$  and  $2 R_\oplus$ . Additionally,  $\eta_{\text{disc}}$  as a function of planet radius  $r$  may be determined by simply evaluating Equation 7 along a grid of SNR distributions corresponding to an evenly spaced grid in planet radius. Figure 5 illustrates this result: discovery completeness rises from zero just below  $r = 0.5 R_\oplus$  to nearly unity at  $r = 4 R_\oplus$ .

### 3.3. Planet Radii

Before being able to apply the formalism discussed in §2.2 to derive an estimate of the planet radius function, we must first understand the radii of the detected planets. The radius of a transiting planet is extracted from its transit light curve, the depth of which reveals—approximately—the planet-star radius ratio  $r_p/R_\star$ . However, there are subtle degeneracies between the radius ratio and other parameters of the fit (impact parameter in particular), especially when the duration of the transit begins to be comparable to the photometric integration time. The *Kepler* long-cadence integration time is 29.4 minutes, and since many of the stars in this particular study are relatively small, the durations of the transits are often only 2-3 times this. Consequently, the shapes of the transits are artificially smoothed, appearing more V-shaped than they would be with shorter-cadence data, and leading to even greater degeneracy between the planet/star radius ratio and the transit impact parameter.

Thus, in order to understand each planet radius, the NExSci archive catalog values (derived from the *Kepler* pipeline maximum-likelihood fits) are insufficient: a Markov Chain Monte Carlo (MCMC) approach is required. Our procedure for extracting, detrending and fitting the transit signals will be described in detail in a forthcoming publication (Swift et al. in prep), but we briefly summarize the process here. Pre-search Data Conditioning Simple Aperture Photometry (PDCSAP) light curves are used from the 23rd data release for our analyses. Transit signals are selected from the light curves based on the catalog values for planet period and duration. Data within 1.5 times the duration of the transit center is preserved and the out of transit data is detrended using a linear model. We then use the `emcee` python module (Foreman-Mackey et al. 2012) to sample the posterior probability distributions for the parameters in our transit model based on Mandel & Agol (2002) using a quadratic stellar limb darkening with coefficients taken from Claret et al. (2012).

To convert the posterior PDFs for  $r_p/R_\star$  of each transit signal to a posterior on  $r_p$ , we do a joint Monte Carlo sampling from the  $r_p/R_\star$  distribution and a distribution for  $R_\star$  according to the derived values from either spectroscopic studies (where available) or Dressing & Charbonneau (2013).

### 3.4. False Positive Probabilities

A transiting planet candidate may be an astrophysical false positive rather than a *bona fide* planet, as has been discussed extensively in literature regarding the *Kepler* survey. In this work, we calculate the false positive probabilities (FPPs) of each of the Cool KOI candidates using the procedure described in Morton (2012), supplemented

by the dispositions on the NExSci archive (for dispositions of “FALSE POSITIVE” we assign FPP=1). Calculating FPP for a transit candidate inevitably requires a prior assumption about planet occurrence; in Morton (2012) this is quantified by the “specific planet occurrence rate”  $f_{p,i}$ , which is an assumed occurrence rate of planets between  $2/3$  and  $4/3$  of the radius of planet  $i$ .

As the goal of the current study is determining exactly the quantity that is used for the planet prior in the FPP calculations (planet occurrence as a function of radius), and since the radius function study itself depends on FPP, there is an opportunity for iterative convergence. In other words, we may start with initial (not necessarily carefully calculated) values for  $f_{p,i}$ , determine the radius distribution  $\hat{\phi}_r^{150}$  (as described in the following subsection) using the FPP values implied by this assumption, then recalculate  $f_{p,i}$  by integrating over the derived radius distribution. If these “post-calculated”  $f_{p,i}$  do not match the initial values used, then the FPPs can be recalculated using these new  $f_{p,i}$  values, and the process repeated until the output  $f_{p,i}$  match the input. The final result of this process is that of the 130 KOIs in this sample, 11 have FPP > 0.9 (and are thus essentially ignored in generating the radius distribution) and 95 have FPP < 0.10. FPPs for individual KOIs are listed in Table 1. We note that these FPP calculations are based on the *Kepler* pipeline-derived  $r/R_\star$  values, so for KOIs in this sample whose MCMC-derived radii are more than 30% discrepant from the *Kepler* pipeline-derived values, we manually set a lower limit to the FPP of 0.1; these KOIs (15) are indicated in Table 1 as well.

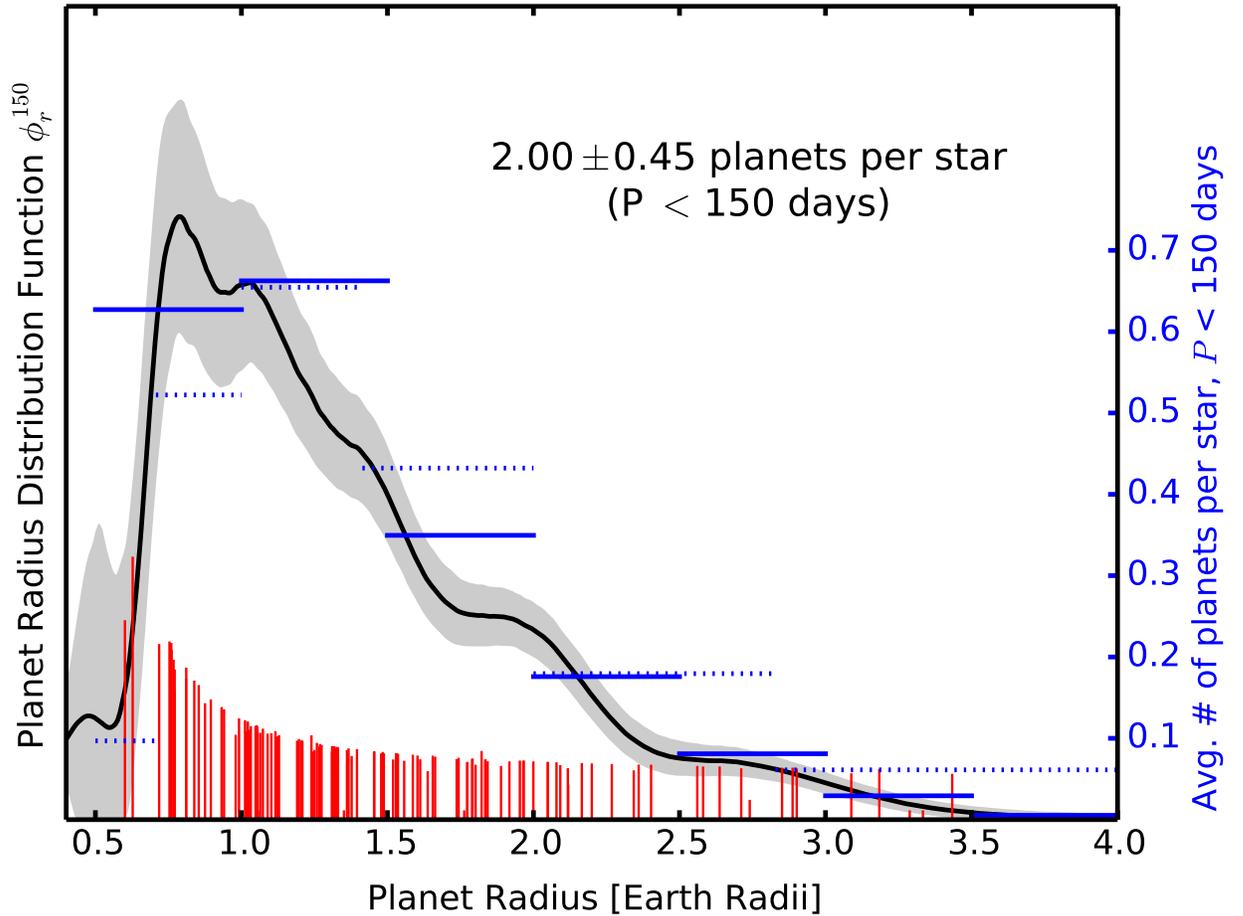
### 3.5. Radius Distribution

With the  $r_p$  posterior distribution  $p_{r,i}(r)$  for each KOI in hand, we then construct the wKDE to estimate the planet radius function. The discovery efficiency  $\eta_{\text{disc},i}$  for each planet is given by integrating the curve in Figure 5 over  $p_{r,i}(r)$ , and multiplied with the average transit probability of 0.026 gives the total efficiency factor  $\eta_i$ . We then calculate FPP values using the iterative method described in §3.4, which combined with  $\eta_i$  gives  $w_i$ , according to Equation 4, and subsequently the radius function  $\hat{\phi}_r^{150}$ . We choose the smoothing kernel for each planet to be  $p_{r,i}(r)$ , without any additional smoothing, and de-bias the estimator and calculate its variance according to the procedure described in §2.2.2. Figure 6 illustrates the result, with the  $1\sigma$  uncertainty region as the grey shaded region. Table 1 lists the planet radii with uncertainties, discovery efficiencies, and FPPs used to generate this radius distribution.

## 4. RESULTS

The overall normalization of the estimated radius function shown in Figure 6 indicates that there are  $2.00 \pm 0.45$  planets per cool star with periods <150d. The most notable feature of this distribution is that it rises more or less smoothly with decreasing radius down to below  $1 R_\oplus$ . The estimator then appears to decrease again for radii smaller than  $\sim 0.8 R_\oplus$ . While this turnover may indeed be real, we show in §5.3 that such a feature may be present even if the true underlying distribution continues to rise.

If the turnover is indeed a robust feature of the distribution, its explanation might be similar to our current



**Figure 6.** The empirical radius distribution of planets orbiting M dwarfs with periods  $< 150$  days (black continuous curve), estimated with a weighted kernel density estimator (wKDE; see §3.5), with the bootstrap resampling-derived  $1\sigma$  uncertainty swath shaded grey—analogueous to a running poisson error bar. The detection efficiency as a function of signal-to-noise ratio has been quantified according to the estimate from the *Kepler* detection pipeline shown in Figure 4. The blue horizontal lines represent the standard “occurrence rate per bin” calculations for this sample, derived by integrating the density estimator over each bin. The solid blue lines are linearly-spaced bins; the dotted blue lines are the logarithmically spaced bins used by Dressing & Charbonneau (2013) and Petigura et al. (2013). The vertical red lines represent the radii of individual planets in the sample, with their heights being proportional to the weight factors  $w_i$ . There is an average of  $2.00 \pm 0.45$  planets per cool star in orbits  $< 150$  days over this radius range, and there is an average of greater than 1 planet per cool star in this period range for radii between  $0.5$  and  $1.5 R_{\oplus}$ .

understanding of the origins of the inner Solar System (Goldreich et al. 2004; Chambers 2001): a large number of isolation-mass protoplanets form quickly, and once the gas and planetesimal disk dissipates, a period of dynamical instability follows, at the end of which typically only a few larger planets remain, the rest having been either destroyed (or merged) via collisions or swallowed by the host star. It is certainly plausible that  $\sim 1 R_{\oplus}$  planets might generically be the most likely outcome of this process, as this is precisely what has happened with the inner Solar System, with an outcome of two planets about the size of Earth.

This distribution also indicates that planets larger than  $\sim 3 R_{\oplus}$  are very rare around cool stars, consistent with the findings of RV surveys (Endl et al. 2003; Johnson et al. 2010b,a; Bonfils et al. 2013). There has been one hot Jupiter identified around a star in this sample (KOI-254b/Kepler-45b Johnson et al. 2012) and another recent discovery of note (Triaud et al. 2013), but such plan-

ets are clearly exceptional—the vast majority of close-in planets around cool stars are smaller than  $\sim 3 R_{\oplus}$ . Even Gliese 1214b (Charbonneau et al. 2009), by far the best-studied planet around an M dwarf to date, appears to be an exception to the typical system, as its radius of  $2.7 R_{\oplus}$  lies far down the tail of this distribution. In fact, there are  $\sim 20\times$  more planets smaller than Gl 1214b than there are larger than Gl 1214b—this bodes very well for the future of ground-based surveys, both transit and RV, as they become more sensitive to smaller planets.

We also note that Figure 6 illustrates how the qualitative interpretation of the radius function changes significantly moving from histogram presentation to the non-parametric density estimator. Overlaid on the the  $\phi_r^{150}$  function plotted in Figure 6 are histograms of planet occurrence rate in two different sets of bins—the solid blue lines being linearly spaced bins and the dotted lines being logarithmically spaced bins, the binning choice for most *Kepler* occurrence analyses to date Howard et al.

**Table 1**  
Data Used in Radius wKDE

KOI	$R_{p,\oplus}$	$+1\sigma$	$-1\sigma$	$\eta_{\text{disc}}$	FPP	$w_i$
2842.03 <sup>a</sup>	0.60	0.24	0.18	0.21	0.1 <sup>c</sup>	162.6
961.03 <sup>a</sup>	0.63	0.16	0.15	0.16	0.1 <sup>c</sup>	214.7
4252.01 <sup>a</sup>	0.72	0.15	0.09	0.26	0.018	143.0
2662.01 <sup>a</sup>	0.75	0.10	0.07	0.26	0.0041	145.1
2842.01 <sup>a</sup>	0.76	0.22	0.18	0.29	0.1 <sup>c</sup>	121.0
961.01 <sup>a</sup>	0.76	0.18	0.18	0.26	0.0091	143.9
1843.02 <sup>a</sup>	0.76	0.10	0.07	0.28	0.01	138.3
2542.01 <sup>a</sup>	0.77	0.13	0.12	0.28	0.069	130.1
255.01 <sup>a</sup>	0.77	0.19	0.12	0.33	0.1 <sup>c</sup>	105.3
4875.01 <sup>a</sup>	0.77	0.17	0.13	0.31	0.021	122.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
251.01 <sup>a</sup>	2.89	0.21	0.22	0.92	0.0017	41.7
250.02 <sup>a</sup>	2.90	0.29	0.34	0.92	0.02	41.0
2793.01 <sup>b</sup>	3.09	0.43	0.38	0.93	0.1 <sup>c</sup>	37.1
250.01 <sup>a</sup>	3.18	0.30	0.30	0.94	0.011	40.4
1006.01 <sup>b</sup>	3.29	0.50	0.39	0.95	1 <sup>c</sup>	0.0
886.01 <sup>a</sup>	3.33	11.09	1.95	0.76	1	0.0
1879.01 <sup>a</sup>	3.43	0.58	0.53	0.95	0.096	36.6
531.01 <sup>a</sup>	4.14	0.47	0.37	0.98	0.99	0.3
1681.01 <sup>a</sup>	4.69	19.40	3.19	0.68	0.99	0.3
2992.01 <sup>b</sup>	5.18	31.57	3.52	0.59	0.88	7.7

<sup>a</sup> Planet radius based on spectroscopic stellar parameters from the analysis of Muirhead et al. (2012a) or Muirhead et al. (in prep).

<sup>b</sup> Spectroscopic stellar characterization not available, so planet radius based on stellar parameters from Dressing & Charbonneau (2013)

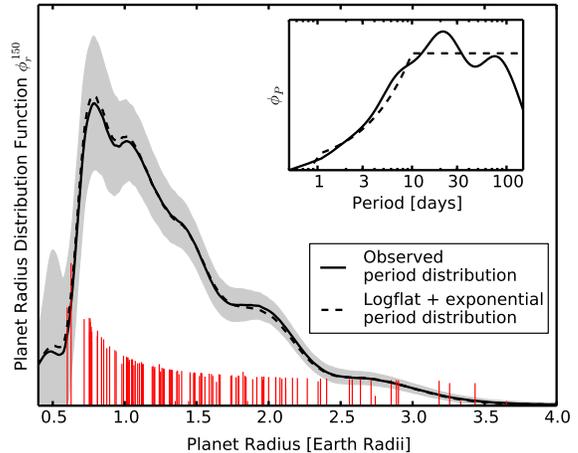
<sup>c</sup> FPPs for these KOIs are not calculated. For those that are dispositioned “FALSE POSITIVE” on the NExScI archive, a value of 1 is assigned; for those for which the *Kepler* pipeline-derived planet radii are more than 30% discrepant from the  $R_p$  listed here, a FPP of 0.10 is assigned.

(2012); Dong & Zhu (2012); Dressing & Charbonneau (2013); Fressin et al. (2013); Petigura & Marcy (2013); Petigura et al. (2013).

The qualitative understanding of planet occurrence patterns communicated by these two different binning schemes is quite different, and both obscure the true underlying detail of the distribution. A quick glance at the occurrence in logarithmic bins gives an impression that the peak of the planet radius distribution is between 1 and 1.5  $R_{\oplus}$ , perhaps around 1.25  $R_{\oplus}$ , while the true peak is really below 1  $R_{\oplus}$ . The linearly binned histogram tells a different story: according to this binning, the impression is that planet occurrence rises with decreasing radius until 1.5  $R_{\oplus}$  and then is constant from there down to 0.5  $R_{\oplus}$ . This characterization is also quite inaccurate, as the planet radius distribution function at 0.8  $R_{\oplus}$  is actually nearly twice that at 1.5  $R_{\oplus}$ . Understanding the shape of the planet radius distribution in detail inaccessible to histogram binning is especially important since it is known that the physical properties of planets likely change dramatically between  $\sim 1$  and  $\sim 2$   $R_{\oplus}$  (Weiss & Marcy 2014; Marcy et al. 2014; Lopez & Fortney 2013).

## 5. DISCUSSION

In this section, we first discuss the assumption we have made that the period distribution of planets is independent of radius. We then explore differences between the methods presented in this paper and those more commonly used in the literature to derive the planet radius distribution. Finally, we validate that the approach used in this work more accurately recovers the true radius



**Figure 7.** Comparing the radius function derived in this work to one derived assuming the period distribution of all planets is as the dashed line in the inset figure, rather than the solid line. The difference in the resulting distribution is negligible; within the uncertainties intrinsic to the de-biasing procedure.

function than does the more commonly used approach.

### 5.1. Period Distribution Assumption

Figure 1 illustrates very clearly why the detected population of small planets in the Cool KOI sample is indeed very likely incomplete, showing that where the detected period distribution of the smallest of the Cool KOIs drops off is right around the periods where the known short-period small KOIs would have become undetectable. This is the motivation behind the period redistribution procedure we use to calculate  $\eta_{\text{disc},i}$  in §3—correcting for the undetectable longer-period small planets. However, the nature of this correction as applied in this work—using the implied all-planet period distribution for each planet—merits some discussion.

There are certainly both physical reasons and observational suggestions to believe that the planet period distribution is *not* completely independent of radius. In particular, Howard et al. (2012) finds (shown in their Figure 6) that the fraction of short-period planets that are large (4-8  $R_{\oplus}$ ) is smaller than the fraction of longer-period planets that are large; in other words, the period distribution of larger planets decreases (heading towards shorter periods) sooner than does the distribution of smaller planets (2-4  $R_{\oplus}$ ). Dong & Zhu (2012) present a similar finding. While there is not yet compelling evidence that this same effect has been detected for planets smaller than 2  $R_{\oplus}$ , simple physical considerations such as increasing stellar insolation (Weiss & Marcy 2013) might reasonably contribute to a dearth of larger planets on short-period orbits. However, as there is no corresponding clear physical explanation for the absence of smaller planets in longer orbits, it is reasonable to assume that they do in fact exist, and that their period distribution might resemble the period distribution of the larger planets that are detected in such orbits.

In order to explore this in detail, we repeat the analysis assuming a modified period distribution, shown in the inset of Figure 7. Rather than using a wKDE built from the observed periods of the detected planets, we use a log-flat distribution for periods greater than 10 days, with

an exponential cutoff short of 10 days. Qualitatively, this seems to approximately match the observed distribution, except that instead of tailing off towards longer periods it remains constant. The radius function that results from assuming this period distribution is plotted as the dashed line in the main panel of Figure 7, and differs only very slightly from the original analysis, within the noise inherent in the de-biasing procedure. The reason for this negligible difference is that the bulk of the correction for small planets happens when it is assumed that the planets observed with  $\lesssim 1$ d periods can also exist at periods of  $\sim 10$ 's of days—in other words, a  $0.7 R_{\oplus}$  planet is just as undetectable at a period of 40 days as at 100 days. We thus conclude that our results are not very sensitive to the details of the longer-period radius distribution, but rather depend only on the assumption that small planets do indeed exist at periods beyond which they are detectable in a manner roughly similar to that of larger planets.

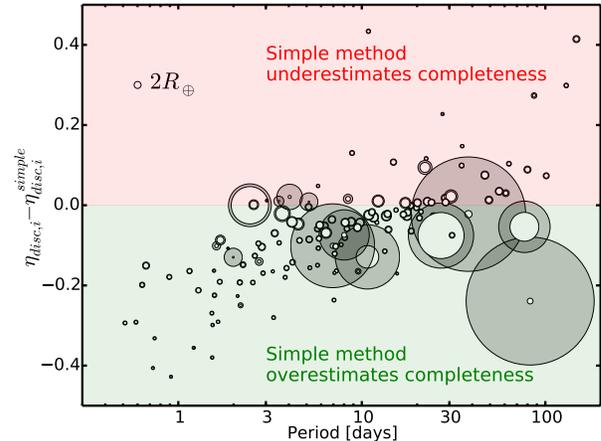
### 5.2. Comparison with other methods

The method we have used to calculate the period-marginalized radius function contains several significant differences from previous planet occurrence rate studies. In particular, we use a survey- and period-averaged approach to calculating both the discovery efficiency of a planet and the transit probability correction factor, as opposed to the more traditionally used method of correcting each planet individually—that is, simulating the planet around other stars at only its observed period, and using each planet’s individual transit probability. We summarize this difference as calculating the radius function while “pre-marginalizing” over period.

Figure 8 illustrates the difference in discovery efficiency calculated using the method used in this work ( $\eta_{\text{disc},i}$  as described in §3) and that which would be calculated by only simulating planets around other stars at the single detected period:  $\eta_{\text{disc},i}^{\text{simple}}$ . In this figure, the size of the points is proportional to planet radius, with the annuli representing the  $\pm 1\sigma$  uncertainties from the radius posteriors. Predictably, planets discovered at short periods—predominantly small—get smaller  $\eta_{\text{disc},i}$  (larger correction factor) than the simple “post-marginalized” calculation; this is because the majority of the period-redistributed simulated planets will be at longer periods than the original, and thus have lower SNR. Planets discovered at longer periods tend to have the opposite effect.

We also note that to get  $\eta_{\text{disc},i}$  in our calculations, we integrate the posterior distribution of planet radius over the efficiency curve in Figure 5; for  $\eta_{\text{disc},i}^{\text{simple}}$ , we simulate the alternative scenarios as having the planet radius fixed to be the median of its posterior—this explains the scatter of this difference around a monotonic relationship to period. Additionally, the efficiency curve we use is not just a simple SNR threshold, as has been used in some previous studies.

In addition, in order to correct for the planets missed by *Kepler* due to random orbital orientations, we do not correct each planet individually for its observed transit probability; rather we use the same single survey- and period-averaged transit probability for every planet. The justification for this is that only way to properly determine  $\eta_i$  is through the full Monte Carlo procedure dis-

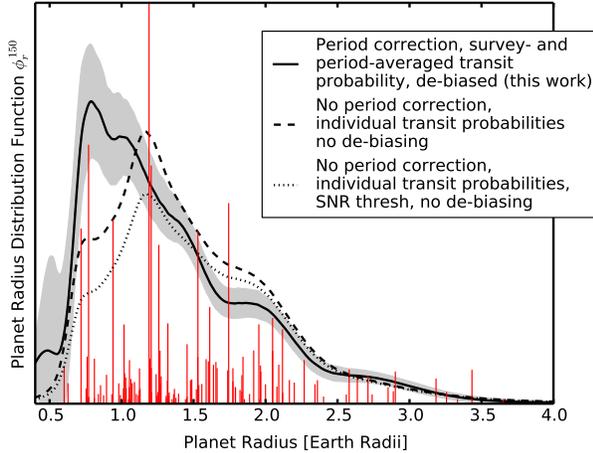


**Figure 8.** The difference between the discovery fraction  $\eta_{\text{disc},i}$  calculated by simulating “alternative scenario” planets around other stars with a distribution of periods, and  $\eta_{\text{disc},i}^{\text{simple}}$ , calculated by simulating planets only at the discovered periods. The sizes of the points are proportional to planet radius, with the annuli representing the  $\pm 1\sigma$  uncertainties from the MCMC posteriors. The discovery fractions of smaller planets, which are found primarily at shorter periods, are typically underestimated by the simple method.

cussed in §3, where planets are simulated isotropically, the SNR distribution of the ones with transiting orientations are used to calculate  $\eta_{\text{disc},i}$ , and the fraction of *the whole simulation* that transits is used as  $\eta_{\text{tr},i}$ . Alternatively, one could imagine ignoring the factorization entirely and just assigning  $\text{SNR} = 0$  (with detectability = 0) to the non-transiting planets in order to directly calculate  $\eta_i$ , and the result would be the same (though less intuitive).

Figure 9 illustrates how the results of this work would differ if we used the same data, but did our calculations according to the more widely used methods of completeness correction. The dotted line illustrates a calculation in the style of Dressing & Charbonneau (2013) or Howard et al. (2012). Each planet is corrected individually for its own transit probability rather than using a global average, and the detection efficiency curve is taken to be a step function at  $\text{SNR} = 7.1$ . We see that the occurrence rate of small planets is significantly underestimated. The dashed line is an improved version of this calculation, using the detection efficiency as a function of SNR from Figure 5, but still correcting for transit geometry using only individual transit probabilities and without period redistribution in the  $\eta_{\text{disc}}$  calculation (this is analogous to the method used by Petigura et al. (2013)). We do not de-bias either the dotted or dashed distributions. The overall normalization is lower than our calculations (solid line) and the qualitative shape of the radius distribution has shifted, showing a peak around  $1.2 R_{\oplus}$  and a clear decline below, rather than the continued rise to below  $1 R_{\oplus}$  that we find.

This qualitative reason for this discrepancy is quite simple: the dashed and dotted lines in Figure 9 are plotted down to below the planet radius where the survey is able to reliably detect planets on 150-day orbits (in this case, about  $1 R_{\oplus}$ ). Whereas the method used in this paper “pre-marginalizes” over period, enabling ex-



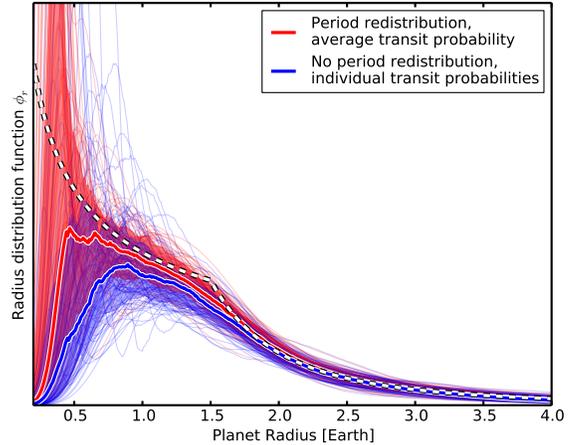
**Figure 9.** Comparing the radius function calculated in this paper (solid line with grey uncertainty region) with the radius distributions that would be derived by calculating completeness corrections according to the methods typically used in the literature. Both the dashed and dotted lines correct for transit probability on an individual basis—i.e., each weight factor  $w_i$  is proportional to the inverse of the geometric transit probability of planet  $i$ , rather than using the survey- and period-averaged transit probability we advocate. The dashed line uses the same discovery efficiency as a function of SNR as we do in this work (Figure 5), whereas the dotted line uses a sharp threshold cutoff at  $\text{SNR} = 7.1$ . The heights of the red lines are proportional to the weights  $w_i$  under the dashed-line calculation. Neither the dashed nor dotted lines implement the de-biasing procedure, which we introduce in this work. We see that using individual transit probabilities moves the peak of the radius function to about  $0.25 R_{\oplus}$  larger than it would otherwise be; in addition, using a SNR threshold significantly decreases the assumed occurrence rate of smaller planets.

trapolation of the smaller-radius population to longer periods, the “simple” individual method does not do this. The simple method implies “post-marginalization,” which can only be valid if the survey is actually sensitive enough to planets in the longest-period and smallest-radius corner of the period-radius space under consideration. So the simple method is not wrong if properly applied, but it cannot be used to estimate the radius function down to as low a radius as the method advocated in this paper. Additionally, it is easy to see from Figure 9 how the shape of the simple-method estimator is much more sensitive to random fluctuations in the data (due to the large correction factors for long-period planets) than the pre-marginalized estimator; this can also cause spurious peaks in the distribution that are strongly dependent on the particular realization of the data that we happen to observe.

### 5.3. Validation of methods, and the limitations of radius function estimation

While we have demonstrated and discussed differences between the planet radius function derived using two different techniques of calculating individual weights, the question still remains whether or how well either can actually accurately reconstruct the true radius distribution.

To explore this, we run a Monte Carlo experiment where we first generate 100 different transit survey data sets with planet frequency and radii drawn from a known distribution—a broken power law normalized to 3 planets/star on the interval  $[0.3, 4.0] R_{\oplus}$ , proportional to  $r_p^{0.5}$



**Figure 10.** A validation test of the wKDE method presented in this work, and comparison of two different completeness-correction recipes. We generate 100 mock transit survey datasets, using the same target stars and signal detection criteria as in this study, with planet radii drawn from the distribution illustrated by the dashed white line. Each thin line is a de-biased wKDE radius function derived according to the methodology presented in this paper. Two different methods are used to calculate the weights for the wKDEs—red lines use the period-averaging strategy we advocate in this paper; blue lines use the more standard procedure of using individually calculated transit probabilities and calculating the discovery efficiency by simulating planets around other stars by fixing the observed period. Heavy solid lines are the median of the ensemble of distributions for each color, and the colored swaths indicate the  $\pm 1\sigma$  percentile range of the thin lines. Both methods recreate the distribution nearly exactly for  $r_p > 1.5 R_{\oplus}$ ; for  $r_p < 1 R_{\oplus}$ , the period-averaging method performs significantly better. Notably, the blue method typically predicts a turnover of the radius distribution around  $1 R_{\oplus}$ , even though the true distribution continues to rise. Conversely, the typical red reconstruction does indeed capture a continued rise down to  $0.5 R_{\oplus}$ , where the detection efficiency is only 5%. However, the variance of this estimator is quite large at these small radii, and many of the data realizations in both methods result in a turnover. Therefore, this experiment demonstrates that detection of a turnover in the reconstructed radius distribution does not rule out a continued rise in the true underlying distribution.

for  $r_p < 1.5 R_{\oplus}$  and to  $r_p^{-3}$  for  $r_p > 1.5 R_{\oplus}$  (the same distribution used in the de-biasing demonstration in Figure 2). The periods of these simulated planets are assigned according to the log-flat/exponential distribution illustrated in Figure 7. For each of these mock data sets, we derive and de-bias the wKDE estimator  $\hat{\phi}_r^{150}$  two different ways: first calculating weights  $w_i$  according to the procedure described in §3, and second calculating weights according to the “simple” prescription, without period redistribution and using individual transit probabilities.

Figure 10 illustrates the results of this experiment. Each thin red line is a de-biased wKDE derived using the methods presented in this paper to calculate detection efficiencies; each thin blue line is a de-biased wKDE using the individual period/individual transit probability (simple) method. The thicker lines and colored bands indicate the respective median and  $\pm 1\sigma$  ranges from this ensemble. The white dashed line is the true underlying distribution from which the planet radii are drawn.

There are several important points this experiment demonstrates. First, both methods correctly recover the underlying distribution nearly perfectly for radii larger than  $1.5 R_{\oplus}$ , validating the accuracy of the technique.

Below  $1.5 R_{\oplus}$ , the two methods begin to diverge, with ours sticking closely to the true distribution until  $r_p \lesssim 0.5 R_{\oplus}$ , but the simple method beginning to significantly underestimate planet occurrence by  $r_p \lesssim 1 R_{\oplus}$ . Indeed, the typical result of the simple method is an estimate of a turnover in the radius function at  $\sim 1 R_{\oplus}$ —a strong qualitative discrepancy from the true distribution.

We also note that at the very low end of the distribution ( $r_p \lesssim 0.5 R_{\oplus}$ ), neither method correctly reconstructs the continued rise, though the period-redistribution method does still keep the true distribution within the  $1\sigma$  percentile range while the simple method does not. This may be understood by realizing that if a very small planet does happen to be detected (which will only occur in rare cases since the detection efficiency is only  $\sim 5\%$  at  $r_p = 0.5 R_{\oplus}$ ), it will necessarily be at a very short orbital period, so the weight factor it receives without period correction will be strongly underestimated. This confirms our qualitative understanding of the difference between the two methods—the post-marginalized simple strategy can only be valid down to a radius where the survey is decently complete at the maximum allowed period (this is apparently somewhere between 1 and  $1.5 R_{\oplus}$ ), whereas the pre-marginalized method enables confident extrapolation to significantly below this point.

Perhaps most dramatically, Figure 10 illustrates that even if the true distribution continues to rise down to arbitrarily small radii, the estimator for its shape will typically turn over significantly above the actual detection limit of the survey, and begin to flatten at even larger radius. In other words, our derived radius distribution from actual *Kepler* data that we show in Figure 6 could very plausibly reflect a true underlying distribution that keeps rising continuously down to below  $0.5 R_{\oplus}$ —*despite the fact that a log-binned histogram of the estimated distribution looks like it turns over below  $1.5 R_{\oplus}$ .*

Finally, we note that the analysis of Petigura et al. (2013) has found, in apparent contradiction with our results, that planets in the  $2.0 - 2.8 R_{\oplus}$  radius bin are more common than planets in the  $1.4 - 2.0 R_{\oplus}$  bin, which are in turn more common than in the  $1.0 - 1.4 R_{\oplus}$  bin. There are several possible explanations for this discrepancy. First, the bins in question are logarithmically spaced, which immediately exaggerates the presence of any turnover, though this alone is not sufficient to explain the difference in results. Second, while the bins used by Petigura et al. (2013) are chosen to all have significant numbers of detected planets, the completeness does vary from  $\sim 70\%$  to  $\sim 10\%$  within the smallest, longest-period bin used for the radius function reconstruction, potentially leaving room for some of the effects that distinguish the blue curve from the red in Figure 10. Another contributing explanation may also be that the only planets counted in Petigura et al. (2013) were the most-detectable planets in each system, which will lead to preferential underestimation of occurrence in the smallest radius bins. Or most simply, this difference may just be a function of the stellar target sample considered—Petigura et al. (2013) used only Solar-like (G/K) stars in their target sample, whereas we use only stars with  $T_{\text{eff}} < 4000$  K. If this is the reason, further study of *Kepler* results should distinguish a radius function that changes shape with increasing host-star tem-

perature.

## 6. CONCLUSIONS

We introduce and validate a simple non-parametric method of analyzing the empirical shape of the period-marginalized planet radius distribution from a transit survey—the weighted kernel density estimator, or wKDE. This estimator is similar to a standard kernel density estimator, except that its overall normalization is constructed to be equal to the total number of planets per star, and that each data point is weighted according to its inverse detection efficiency. While the naive construction of this estimator is most generally biased, we present a bootstrap-based method for de-biasing it, adapted from the “smoothed bootstrap” presented in Narsky & Porter (2013). We also show that the detection efficiency is best computed with a “pre-marginalization” procedure: re-distributing planets at all possible periods when calculating how many target stars around which a particular planet could have been observed, and using a survey- and period-averaged transit probability. Additionally, we demonstrate that it is important to use a realistic detection efficiency as a rising function of signal-to-noise ratio, rather than a strict cutoff, to correctly calculate the occurrence rates of smaller planets. We also emphasize that presentation of the planet radius function in histograms—especially with logarithmically spaced bins—runs a risk of qualitatively misrepresenting the true shape of the distribution.

Applying this analysis to the 130 planet candidates in the Q1-Q12 KOI catalog with periods less than 150 days discovered around the cool ( $T_{\text{eff}} < 4000$  K) *Kepler* targets photometrically characterized by Dressing & Charbonneau (2013), incorporating their individually calculated false positive probabilities following Morton (2012) as well as new MCMC fits for planet radii (Swift et al., in prep), we find that the planet distribution continues to rise continuously down to at least  $\sim 1 R_{\oplus}$  and possibly below. We detect a possible turnover in the radius distribution below  $\sim 0.8 R_{\oplus}$  but we demonstrate that it is very plausible that this may reflect an underlying distribution that continues to rise.

This is the first radius-function reconstruction study that is sensitive to this range of planetary radius, and it appears to contradict the results of Petigura et al. (2013), who use a “post-marginalization” completeness correction method and see a peak in the radius function around  $\sim 2 R_{\oplus}$ . This contradiction may be due to either a confluence of small factors that could add up to a qualitative miscalculation in the Petigura et al. (2013) study, or it may simply be due to the different host star populations considered (G/K in that study, as opposed to M stars in this). This certainly warrants further investigation, in particular investigating how the radius function may change with stellar type. New detections by future transit missions such as K2, TESS, and PLATO will also provide additional insight into these questions.

If the flattening/turnover of the radius function just below  $1 R_{\oplus}$  is indeed a true feature of the distribution, it invites theoretical exploration, as it would suggest that planets around the size of Earth are the most common to survive the process of system formation and evolution around cool stars. This outcome is certainly plausible, given the observed architecture of our Solar System. And

as a final note, the occurrence pattern of planets around cool stars indicates that there are indeed many planets just beyond the detection threshold of ground-based surveys, as planets larger than Gl 1214b ( $2.7 R_{\oplus}$ ) are  $\sim 20\times$  rarer than planets with  $R_p < 2.7R_{\oplus}$ .

Looking at the occurrence rates of the smallest planets in particular, we may compare with previous studies to estimate the degree to which the estimates of “habitable-zone” planets might change with our improved calculations. This may be accomplished by recognizing that the analysis of Dressing & Charbonneau (2013) used both the post-marginalized method of completeness correction and a SNR=7.1 detection threshold (a calculation in the style of the dotted line in Figure 9. As the integral of our reconstructed radius function on the interval  $[0, 1.4] R_{\oplus}$  is about  $1.6\times$  larger than the integral over the same range of the post-marginalized method using the SNR threshold, we conclude that there should be closer to 0.25 habitable-zone Earth-like planets per cool star, rather than the  $\sim 0.15$  estimated by that work. And if this same correction is made to the calculations of Kopparapu (2013), which use updated HZ calculations but the same occurrence formalism as Dressing & Charbonneau (2013), than this number would become closer to  $\sim 0.8$  planet per star. It is likely that habitable-zone, Earth-sized planets abound throughout the Galaxy in numbers even larger than previously estimated.

Finally, we emphasize that this calculation is based on a target sample of only about 3900 cool stars and a KOI search only through Q12 data. Future pipeline searches and continued observations of cool stars by the K2 mission, as well as future surveys such as TESS and PLATO, will increase this sample size, allowing for strengthened conclusions from the small-planet radius distribution and giving a greater handle on the formation processes of planetary systems around the most numerous stars in the Galaxy. In addition, careful application of these same principles to the entire *Kepler* dataset, as permitted by accurate knowledge of stellar parameters, will continue to uncover important clues to the formation and evolution of all types of planetary systems.

The authors acknowledge John Johnson and the Caltech Exolab for nurturing the intellectual environment in which this work took shape, in particular, conversations with Phil Muirhead, Leslie Rogers, Avi Shporer, Ben Montet, and Jean-Michel Desert. TDM thanks Ed Turner and David Hogg for supportive conversations regarding this analysis, and Frank Porter for very helpful tips regarding density estimation, and for his willingness to share the draft version of his book (Narsky & Porter 2013), by which the bootstrap techniques used in this work were inspired. We also thank Peter Goldreich for an enlightening conversation about rocky planet formation, Dave Charbonneau and Courtney Dressing for helpful comments on an early draft. TDM acknowledges support from the *Kepler* Participating Scientist Program, NASA NNX11AG85G.

#### REFERENCES

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2012, ArXiv e-prints, arXiv:1202.5852

Bonfils, X., Delfosse, X., Udry, S., et al. 2013, A&A, 549, A109

Borucki, W. J., Koch, D. G., Basri, G., et al. 2011, ArXiv e-prints, arXiv:1102.0541

Burke, C. 2013, in prep

Chambers, J. E. 2001, Icarus, 152, 205

Charbonneau, D., Berta, Z. K., Irwin, J., et al. 2009, Nature, 462, 891

Chiang, E., & Laughlin, G. 2012, ArXiv e-prints, arXiv:1211.1673

Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, PASP, 124, 1279

Claret, A., Hauschildt, P. H., & Witte, S. 2012, A&A, 546, A14

Dong, S., & Zhu, Z. 2012, ArXiv e-prints, arXiv:1212.4853

Dressing, C. D., & Charbonneau, D. 2013, ArXiv e-prints, arXiv:1302.1647

Endl, M., Cochran, W. D., Tull, R. G., & MacQueen, P. J. 2003, AJ, 126, 3099

Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2012, ArXiv e-prints, arXiv:1202.6328

Fang, J., & Margot, J.-L. 2012, ApJS, 761, 92

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2012, ArXiv e-prints, arXiv:1202.3665

Fressin, F., Torres, G., Charbonneau, D., et al. 2013, ArXiv e-prints, arXiv:1301.0842

Gaidos, E., & Mann, A. W. 2013, ApJ, 762, 41

Gaudi, B. S. 2005, ApJ, 628, L73

Gaudi, B. S., Seager, S., & Mallen-Ornelas, G. 2005, ApJ, 623, 472

Goldreich, P., Lithwick, Y., & Sari, R. 2004, ARA&A, 42, 549

Hansen, B., & Murray, N. 2013, ArXiv e-prints, arXiv:1301.7431

Hansen, B. M. S., & Murray, N. 2012, ApJ, 751, 158

Howard, A. W., Marcy, G. W., Johnson, J. A., et al. 2010, Science, 330, 653

Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15

Johnson, J. A., Aller, K. M., Howard, A. W., & Crepp, J. R. 2010a, PASP, 122, 905

Johnson, J. A., Howard, A. W., Marcy, G. W., et al. 2010b, PASP, 122, 149

Johnson, J. A., Gazak, J. Z., Apps, K., et al. 2012, AJ, 143, 111

Kopparapu, R. 2013, arxiv:1303.2649

Lopez, E. D., & Fortney, J. J. 2013, ArXiv e-prints, arXiv:1311.0329

Lopez, E. D., Fortney, J. J., & Miller, N. 2012, ApJS, 761, 59

Mandel, K., & Agol, E. 2002, ApJ, 580, L171

Mann, A. W., Gaidos, E., Lépine, S., & Hilton, E. J. 2012, ApJ, 753, 90

Marcy, G. W., & Butler, R. P. 1996, ApJ, 464, L147+

Marcy, G. W., Isaacson, H., Howard, A. W., et al. 2014, ApJS, 210, 20

Mayor, M., & Queloz, D. 1995, Nature, 378, 355

Morton, T. D. 2012, ApJ, 761, 6

Muirhead, P. S., Hamren, K., Schlawin, E., et al. 2012a, ApJ, 750, L37

Muirhead, P. S., Johnson, J. A., Apps, K., et al. 2012b, ApJ, 747, 144

Narsky, I., & Porter, F. C. 2013, Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning (Wiley)

Owen, J. E., & Wu, Y. 2013, ApJ, 775, 105

Petigura, E., & Marcy, G. 2013, in prep

Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, Proceedings of the National Academy of Science, 110, 19273

Raymond, S. N., & Cossou, C. 2014, MNRAS, 440, L11

Swift, J. J., Johnson, J. A., Morton, T. D., et al. 2013, ApJS, 764, 105

Triana, A. H. M. J., Anderson, D. R., Collier Cameron, A., et al. 2013, A&A, 551, A80

Wang, B., & Wang, X. 2007, ArXiv e-prints, http://arxiv.org/abs/0709.1616, arXiv:0709.1616

Weiss, L., & Marcy, G. 2013, in prep

Weiss, L. M., & Marcy, G. W. 2014, ApJ, 783, L6

Wolszczan, A., & Frail, D. A. 1992, Nature, 355, 145

Wright, J. T., Marcy, G. W., Howard, A. W., et al. 2012, ArXiv e-prints, arXiv:1205.2273

Wu, Y., & Lithwick, Y. 2013, ApJ, 772, 74

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