A STATISTICAL RECONSTRUCTION OF THE PLANET POPULATION AROUND KEPLER SOLAR-TYPE STARS

ARI SILBURT1, ERIC GAIDOS2, AND YANQIN WU1

1 Department of Astronomy and Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
2 Department of Geology & Geophysics, University of Hawai‘i at Mānoa, Honolulu, HI 96822, USA

Received 2014 June 23; accepted 2014 November 24; published 2015 January 28

ABSTRACT

Using the cumulative catalog of planets detected by the NASA Kepler mission, we reconstruct the intrinsic occurrence of Earth- to Neptune-size (1–4 $R_\oplus$) planets and their distributions with radius and orbital period. We analyze 76,711 solar-type (0.8 $< R_*/R_\odot < 1.2$) stars with 430 planets on 20–200 day orbits, excluding close-in planets that may have been affected by the proximity to the host star. Our analysis considers errors in planet radii and includes an “iterative simulation” technique that does not bin the data. We find a radius distribution that peaks at 2–2.8 Earth radii, with lower numbers of smaller and larger planets. These planets are uniformly distributed with logarithmic period, and the mean number of such planets per star is 0.46 ± 0.03. The occurrence is ∼0.66 if planets interior to 20 days are included. We estimate the occurrence of Earth-size planets in the “habitable zone” (defined as 1–2 $R_\oplus$, 0.99–1.7 AU for solar-twin stars) as 6.4$^{+3.4}_{-1.4}$%. Our results largely agree with those of Petigura et al., although we find a higher occurrence of 2.8–4 Earth-radii planets. The reasons for this excess are the inclusion of errors in planet radius, updated Huber et al. stellar parameters, and also the exclusion of planets that may have been affected by proximity to the host star.

Key words: astrobiology – methods: statistical – planetary systems

Supporting material: machine-readable table

1. INTRODUCTION

Anaximander of Miletus proposed that there were many Earth-like worlds. In the 25 centuries since the Ionian philosopher’s speculation, we have seen an accelerating convergence toward an answer: many if not most stars have planets, Earth-sized (and presumably rocky) bodies are more common than Jupiter-sized gaseous bodies, and some Earth-sized planets are on orbits in the so-called “habitable zones” of their stars (Schneider et al. 2011). The most recent advances in this field have come from data collected by the Kepler transiting-planet mission (Borucki et al. 2010). Four years of observations of a ∼100 deg2 field have led to the identification thus far of 4234 confirmed or candidate transiting planets, about 84% of the candidates appear to be smaller than Neptune (3.88 $R_\oplus$) (NASA Exoplanet Archive4).

Beyond simple human curiosity about the prevalence of Earth-like planets and possibility of life elsewhere, planet statistics provide an important test of planet formation models (e.g., Benz et al. 2014). For example, the distribution of planets with respect to orbital period and mean motion resonances test models of planet formation and early migration (e.g., Hansen & Murray 2013; Baruteau et al. 2013). Planet radius and period distributions can be combined to reconstruct the distribution of solid mass in disks (e.g., Chiang & Laughlin 2013; Raymond & Cossou 2014).

There have been many previous studies that use the Kepler data to infer the intrinsic population of planets around Kepler target stars, or subsets of those stars (e.g., Catanzarite & Shao 2011; Youdin 2011; Traub 2012; Howard et al. 2012; Fressin et al. 2013; Dressing & Charbonneau 2013; Gaidos 2013; Dong & Zhu 2013; Kopparapu 2013; Petigura et al. 2013). These works differ in their samples and methods, but are broadly consistent in estimating that the occurrence of planets per star on orbits of days to months to be of order unity. Some of these works have estimated $\eta_{R/\oplus}$, the occurrence of Earth-size planets in a circumstellar “habitable zone” (where stellar irradiation is similar to the solar constant), and find that $\eta_{R/\oplus}$ is of an order of tens of percent. Of particular interest to us is the work of Petigura et al. (2013, hereafter PHM13) who performed an independent analysis of the Kepler photometric light curves by identifying candidate planet transits and determining the detection efficiency by injecting synthetic transit signals into Kepler light curves and recovering them. Among the salient conclusions from PHM13 are that the distribution of planets peaks at a radius of 2–2.8 $R_\oplus$ and that $\eta_{R/\oplus}$ is about 22% (using a liberal definition of the habitable zone).

We identify two reasons to revisit the derivation of the Kepler planet population. First, we want to more fully account for the uncertainties and biases in the Kepler data and related observations of the target stars. Second, we wish to consider a “primordial” planet population and restrict our analysis to planets far enough from their host stars such that their properties have not been altered by proximity since their formation.

Precise statements on the occurrence of planets requires rigorous statistical methods, full accounting of errors, and adequate assessment of potential biases. First, while the overall rate of “false positives” among Kepler candidate planets appears to be low (Morton & Johnson 2011; Fressin et al. 2013), it is not uniform across all periods and all sizes (Santerne et al. 2013). Second, determination of planet occurrence from transit surveys requires accurate estimates of detection efficiency (also known as “completeness”), which depends on the parameters (i.e., density and/or radius) of not only the planet host stars but of the entire target catalog. The parameters of Kepler stars were first determined by combining multi-wavelength photometry, stellar models, and Bayesian inference (Brown et al. 2011). Colors of solar-type stars depend only weakly on gravity and metallicity, and parameter values based on photometry have large random

3 Dissertation on the Philosophy of Aristotle, in which his Principal Physical and Metaphysical Dogmas are Unfolded, transl. by T. Taylor, London, 1812.
4 http://exoplanetarchive.ipac.caltech.edu
and systematic errors in effective temperature (Pinsonneault et al. 2012), gravity, and luminosity class (Mann et al. 2012). Spectroscopy, asteroseismology, and improved stellar models have yielded more reliable parameters (Huber et al. 2014), especially for Kepler planet-hosting stars (Kepler Objects of Interest or KOIs). Nevertheless, 70% of all stars in the Huber et al. (2014) catalog have assigned parameters based on Kepler Input Catalog (KIC) photometry; the median upper and lower fractional errors in stellar radius among these solar-type stars is 40% and 10%, respectively. Because the estimated radius of a planet detected by transit depends on the stellar radius, and the probability of transit depends on stellar density, these errors need to be taken into account when computing occurrence rates. Errors in luminosity also affect the certainty with which a planet can be assigned to a habitable zone described by a range of stellar irradiance (Gaidos 2013; Mann et al. 2013).

Third, biases, if uncorrected or unaccounted for, will distort our perspective on planet populations. Mann et al. (2012) found that up to 96% of the reddest Kepler target stars are giants, even though virtually all red KOI hosts are dwarfs for the simple reason that it is extremely difficult to detect a transiting planet around a giant star. They showed that dilution of the target catalog by giants had led to an underestimate of the occurrence rate and an incorrect claim that M dwarf hosts of detected planets are redder and thus more metal-rich than those without detected planets.

Gaidos & Mann (2013) showed that because the Kepler target catalog is essentially magnitude-limited, Malmquist bias combined with uncertainties in stellar parameters means that stellar distances are underestimated and many stars are likely to be more luminous, evolved, and larger than their nominal values. Follow-up observations and analysis thus far seem to confirm this (e.g., Bastien et al. 2014; Everett et al. 2013; Verner et al. 2011). Because transiting planet radius scales increases with stellar radius, planet radii are thus underestimated. Moreover, the rate of planet detection decreases with increasing stellar radius or density, thus for a given planet radius, the detection rate is overestimated and thus the occurrence is underestimated. Detection bias again means that any estimate of this effect based on the host stars of transiting planets is an underestimate: the effect will be greater among stars in the overall target catalog. Another bias is Eddington bias: scatter by error from more populated regions of parameter space into less populated regions will produces the opposite effect, i.e., occurrence will be overestimated (Gaidos & Mann 2013).

Previous analyses of the Kepler planet population have sometimes not taken these errors or biases into account, and instead have considered only Poisson (counting) statistics (e.g., Petigura et al. 2013; Howard et al. 2012). Finally, many analyses were performed by binning the data into discrete bins of planet radius $R_p$ and orbital period $P$. While simple and readily explicable, the binning method runs the risk of masking details of a distribution, especially that of radius, which may be important for testing theoretical models.

In this work, we consider a “primordial” population of planets, as opposed to one that has evolved under the influence of the host star. Effects of the latter, including tidal heating (Jackson et al. 2008), atmospheric escape (Tian et al. 2005), Ohmic heating (Batygin et al. 2011), and impact erosion (Marcus et al. 2009), act with an efficiency that is inversely proportional to the distance to the host star. In particular, Owen & Wu (2013) proposed that photoevaporation by stellar X-ray and ultraviolet irradiation have effectively removed the hydrogen envelopes of close-in planets ($P \leq 10$ days), leading to the observed paucity of super-Earth sized planets in that neighborhood. This process was also investigated for a few Kepler systems by Lopez et al. (2012). Regardless of the mechanism, the distinctiveness of the $P < 20$ day and $P > 20$ day populations (see, e.g., Youdin 2011) suggests that any analysis treat these separately. In this study, we focus exclusively on the latter population as we believe it is more likely to represent the “primordial” state. On the other hand, because of Kepler’s low efficiency at detecting long-period planets (see Section 3.1), we are forced to limit our consideration to planets with $P < 200$ days.

Practical reasons also limit the range of planet radius $R_p$ considered. Although Kepler can readily detect a transiting giant planet, the occurrence of these objects is indubitably much lower than that of smaller planets. The distribution with planet radius falls to a very low level beyond Neptune-size objects: only 8% of Kepler candidate planets have nominal radii $> 8 R_{\oplus}$, and the false-positive rate increases as well (Santerne et al. 2012; Colón et al. 2012). Conversely, Kepler can detect planets smaller than 1 $R_{\oplus}$ for only a tiny fraction of stars, mostly M dwarfs. For these reasons we restrict our analysis to a radius range of $1 - 4 R_{\oplus}$ over which statistically rigorous analyses can be performed.

In this contribution, we infer the intrinsic distribution of planets with $20 < P < 200$ days (equivalent to 0.16–0.67 AU) and $1 R_{\oplus} < R_p < 4 R_{\oplus}$ around solar-type stars as observed by Kepler over its entire mission (Quarters 1–16). This analysis includes the effects of errors in stellar and planet radius, and takes into account some of the biases that may affect previous works. We introduce a method of iterative simulation (IS) to determine the radius distribution of planets without resorting to binning. We compare our results with those of PHM13 and also carry out a detailed comparison of the two methods to understand the source of any discrepancies. Finally, we use our simulations to assess the effect of systematic bias, namely an overall underestimate of stellar radius, on inferences of a planet population from the Kepler catalog.

2. METHODS

2.1. Catalogs of Stars and Planets

We construct a stellar sample from the Huber et al. (2014) catalog of 196,468 stars observed during the Kepler mission (Quarters 1–16), selecting “Sun-like” stars with radii $0.8 R_{\odot} < R_s < 1.2 R_{\odot}$. We restrict the sample to stars with a Kepler magnitude $K_p < 15.5$ to avoid faint stars with uncertain properties and noisy light curves. This leaves a sample of 76,711 solar-type stars, hereafter known as the “Solar76k” sample. Our planet sample is constructed from the 2014 February 26 version of the KOI catalog (Ramirez et al. 2014). Planet and stellar parameters are updated with values from Huber et al. (2014) where available. In addition to the cuts made on our stellar sample we also require $20 < P < 200$ days, $0.5 R_{\oplus} < R_p < 6 R_{\oplus}$ and $S/N > 12$. Although we are only interested in the occurrence of planets $1 R_{\oplus} < R_p < 4 R_{\oplus}$, we use a larger radius range for our analysis since these planets have a non-zero probability of being in our region of interest after accounting for radius errors (Section 2.3). This leaves us with 430 candidate planets, hereafter known as the “430KOI” sample. The full 430KOI data set is included in Table 4.

We also construct a second planet sample (used only in Section 3.2) using the same cuts above except relaxing the period restriction to $5 < P < 200$. This leaves us with 1052 KOIs, hereafter known as the “1052KOI” sample.
To ease comparison with PHM13, we retrieve their vetted sample of 603 planet candidates that fall within 5 < \( P < 100 \) days. We update the stellar parameters where possible, using Huber et al. (2014). We hereafter refer to this planet sample as the “603PHM” sample.

### 2.2. Simulated Planet Detections

Our simulator synthesizes single planet–star pairs,\(^5\) drawing stars from one of the catalogs described above, and planet parameters from a large “master” population as we describe in Section 2.3. We calculate whether each planet transits its host star in a probabilistic manner, and then determine whether Kepler could have detected it. We compare the properties of these simulated detections with the observed candidate Kepler planets, and modify the master population using two different techniques: iterative Monte Carlo Markov Chain (MCMC) and IS. In implementing our simulations we make two critical assumptions. First, that the orbital periods and radii of Kepler planets (as well as orbital eccentricities), are independently distributed, i.e., the occurrence is a separable function of period and radius:

\[
\frac{d^2N}{d\log P \ d\log R} = p(P) \nu(R),
\]

where \( p \) and \( r \) are some yet-to-be-determined functions, and \( N \) is the total number of planets. Second, we assume that these distributions do not vary over the range of stellar parameters considered. We discuss these assumptions in Section 5.1. In this study, we further specify that the period distribution is a power law:

\[
\frac{dN}{d\log P} = C P^\alpha,
\]

where \( C \) is a normalization constant.

The geometric probability that a planet transits its host star is (Winn 2010),

\[
p = \frac{R_\ast \left( 1 + e \sin \omega \right)}{a \left( 1 - e^2 \right)},
\]

where \( a \) is the semimajor axis, \( e \) the orbital eccentricity, and \( \omega \) the argument of periastron. While \( a \) can be calculated from \( P \) and the estimated mass of the host star, the orbital eccentricity of Kepler planets are unknown and must be estimated statistically. Assuming a Rayleigh distribution with dispersion \( \sigma_e \), Moorhead et al. (2011) estimated \( \sigma_e = 0.2 \) by studying the distribution of transit durations. This is likely affected by uncertain stellar radii and may be an overestimate. Transit timing variable (TTV) studies have led to much smaller eccentricity dispersion (\( \sigma_e \sim a \) few percent), at least in multiple planet systems (Wu & Lithwick 2013; Hadden & Lithwick 2014). Here, we choose \( \sigma_e = 0.18 \) and show that our results are not sensitive to the exact value of \( \sigma_e \) (Section 5.1). The underlying distribution of \( \omega \) can be safely assumed to be uniform over \([0, 2\pi]\). Integrating \( p \) over the distributions of \( e \) and \( \omega \), Equation (3) becomes \( p = p_0 R_\ast / a \), with \( p_0 = 1.073 \). As in previous works, we require that at least three transits have been observed. We scale every transit probability by \( (1/p_0)a/R_\ast \text{max} \), the inverse of the max transit probability. Since planets transit for only \(<5\) deg of inclination, this scaling is used to speed up the rate of the simulation.

We now proceed to assign a transit duration, \( T \), to a given transiting planet. We follow the procedure of Gaidos (2013) by setting

\[
T = \tau^{2/3} P^{1/3} \Delta,
\]

where \( \tau = 2\sqrt{R_\ast^3/(\pi G M_\ast)} \) is the stellar free-fall time, \( G \) the gravitational constant, \( M_\ast \) the stellar mass, and

\[
\Delta = \sqrt{(1 - e^2)(1 - b^2)} \left( 1 + e \cos \omega \right),
\]

with \( b \) being the impact parameter. For \( a \gg R_\ast \), the impact parameter \( b \) is uniformly distributed in the range [0,1]. We then calculate \( dN / d\Delta \), the likelihood of drawing a given \( \Delta \), or rather, its cumulative distribution, \( \tilde{N}(\Delta) \equiv \int_0^\Delta dN / d\Delta \) d\( \Delta \), with the overbar indicating marginalization over \( e \) and \( \omega \). Using the chain rule, we find

\[
\frac{dN}{d\Delta} \frac{d\Delta}{d\Delta} = \frac{dN}{d\alpha} \frac{d\alpha}{d\Delta} = \frac{db}{d\Delta},
\]

where we have used the fact that \( dN / db = 1 \) (i.e., \( b \) is uniformly distributed) for transiting systems. As a result, \( \tilde{N}(\Delta) = \bar{\eta} = \int_0^{\tau} (db / d\Delta) d\Delta \). Inverting Equation (5) then yields

\[
\tilde{N}(\Delta) = \int_0^\Delta \bar{\eta}(e) \ d e \int_0^{2\pi} \sqrt{1 - \frac{\Delta^2(1 + e \cos \omega)^2}{(1 - e^2)}} \ d \omega,
\]

where \( \bar{\eta}(e) \) is the assumed eccentricity distribution.

We also need to assign a radius to each trial planet: this process differs between our MCMC and IS methods and is described in their respective sections. Moreover, in comparing the radius distribution of trial planets to the observations, we must take into account significant uncertainties in the radius of KOIs. We describe how we do this in the next section.

Our detection criterion is based on a comparison between the transit signal, \((R_p/R_\ast)^2\), and the effective noise over the transit duration. Fressin et al. (2013) established that at signal-to-noise \( S/N > 12 \) the false-positive rate among KOIs is very low. PHM13 used this criterion for their analysis and we follow suit. Noise in Kepler light curves is derived from photon (shot) noise, measurement error (e.g., pointing error and instrument noise), and stellar variability (Koch et al. 2010). The Kepler team encapsulates the total noise of each star into quarterly transit observations (Christiansen et al. 2012) values. For a given star in a given quarter, we generate the appropriate noise for transit duration \( T \), by interpolating among the various CDPP values using a power-law relation. Because sources of noise (e.g., stellar variability) are not necessarily “white,” the power-law index can and often does depart from −0.5, the white noise value.

We then calculate the total \( S/N \) of a model star–planet pair as

\[
\frac{S}{N} = \left( \frac{R_p}{R_\ast} \right)^2 \left[ \sum_{j=1}^{16} \frac{n_j}{\text{CDPP}_j} \right]^{1/2},
\]

where \( n_j \) is the number of transits in quarters \( j \), and \( \text{CDPP}_j \) is the interpolated CDPP value for that quarter. \( n_j \) is found by first assigning a randomly drawn phase for each planet, and then counting the number of transits in each quarter. The system is proclaimed detectable if \( S/N > 12 \). The \( S/N \) threshold does not account for noise that is non-Gaussian or non-stationary on timescale shorter than one observing quarter (90 days). However, the conservative requirement that \( S/N > 12 \) for detection partially addresses this limitation and we consider.

\(^5\) We ignore the occurrence of multiple systems and assume that each planet has an independent occurrence.
the possible impact of this simplification in Section 4 when we compare our analysis to PHM13. Other non-stationary effects unaccounted for in Equation (8) include thermal settling events, sudden pixel sensitivity drop offs, and cosmic rays. Equation (8) also does not account for gaps in the data.

2.3. Uncertainties in Planet Radii

As described in Section 1, there are significant uncertainties in the radii of most KOIs (median uncertainty \(= 33\%\), primarily due to our limited knowledge of the host star. For example, this means that there is a non-negligible chance that a planet with a cataloged radius value of \(R_p = 2.5 \ R_\oplus\) is actually Earth-sized or Neptune-sized. To clarify, we are not addressing the issue that some dwarf stars are actually giants (which is addressed in Section 3.3 when we consider that all stars are 25\% larger). Instead, we are assuming that all claimed dwarf stars are truly dwarfs, and are accounting for the fact that the exact radii of these stars are still uncertain (on average) to \(\sim 30\%\).

It is important that uncertainties of such magnitude be considered, and we do this by replacing each nominal radius by a distribution of radii governed by Bayesian statistics. The probability that a planet with a reported radius \(R\) actually has a true radius \(R'\) is given by \(p(R'|R) = q(R|R') r(R')\), where \(r(R')\) is a normalized prior and is the probability that a planet of radius \(R'\) (with same period \(P\)) would be detected by Kepler around a given star. Put another way, \(r(R')\) is essentially the survey completeness (Section 3.1) of planet \(R'\) (having period \(P\)) with respect to the entire Solar76k catalog.

In our treatment, we assume that errors in \(R_\star\) and hence \(R_p\) are normally distributed. This means that \(q(R|R') = q(R'|R)\) because the Gaussian only depends on the square of the difference \(R - R'\). We also assume that errors in stellar radius are uncorrelated. This latter assumption means that a planet with a radius that has been over/underestimated would, on average, produce a weaker/stronger transit signal among the ensemble of target stars and that such a planet would become less/more detectable. If errors in \(R_\star\) were exactly correlated, then errors in \(R_p\) would be unaffected by considerations of detection; if all stars are smaller then their planets will also be smaller by the same proportion, and thus produce transit signals of the same depth.

Provided these assumptions hold, a planet cannot be arbitrarily small, even if the errors in radius are large because it would never have been detected in the first place. The \(r(R')\) factor accounts for this fact. Our prescription for handling radius errors also accounts for the fact that the cataloged radius is more likely to be an underestimate, rather than an overestimate, of the true radius. This effect becomes more pronounced among KOIs with small cataloged radii and large uncertainty. For these cases the result is an error distribution that is no longer a Gaussian but is strongly asymmetric, with a cutoff just below the cataloged radius and an extended tail to larger radii. Example probability distributions in radius for KOI-1338.01 (solid) and KOI-1925.01 (dotted) are shown in Figure 1. The vertical red lines represent the cataloged radius values. As seen in Figure 1, the most likely radius differs from the catalog value.

We implement radius errors into our analysis by replacing each KOI with a probability distribution function (PDF) that is the product of a Gaussian times a prior detection function, which is the fraction of stars around which the planet would be detected (i.e., completeness). We represent the PDF by a large number of Monte Carlo planets drawn from a Gaussian distribution with mean equal to the nominal value of \(R_p\) and standard deviation equal to the cataloged error. We calculate the fraction of stars \(F\) around which each Monte Carlo planet could be detected.

We then compute a normalized cumulative distribution function (CDF) of \(F\) with \(R_p\) for each Monte Carlo set. We can then draw a radius value from each corrected error distribution by comparing the CDF to a uniform random deviate. We create a “master” radius distribution by randomly drawing two million values from all of these distributions according to their CDFs. We use this distribution to represent the inferred radius distribution of observed candidate planets after errors have been accounted for.

2.4. Monte Carlo Markov Chain

We implement the MCMC according to the algorithm by Gelman & Rubin (1992). We discretize the radius–period plane into 16 bins, 4 period bins equally spaced in log \(P\) \((P = [20–40, 40–80, 80–160, 160–200] \text{ days})\), and 4 radius bins equally spaced in log \(R\) \((1–1.4 \ R_\oplus, 1.4–2 \ R_\oplus, 2–2.8 \ R_\oplus, 2.8–4 \ R_\oplus)\). We parameterize the planet population by a set of four parameters: \(\alpha\) (from Equation (2)), and \(\kappa_1, \kappa_2, \kappa_3\), where the latter three parameters are the relative numbers of planets in the first three radius bins to the last bin. For each set of parameters, we generate a mock catalog by simulating 10^5 transiting pairs around a given stellar sample (Section 2.2) while properly taking into account errors in planet radius (Section 2.3). We then compare our mock population against the 430KOI sample by first binning the KOIs in the same manner and then removing a portion from each bin to account for false positives (Table 1 from Fressin et al. 2013).

We then scale our mock catalog down from 10^5 to match the total number of remaining KOIs. The goodness of fit is measured by comparing the simulated number of planets in each of the 16 bins, \(S_i\), versus that of the observed, \(D_i\),

\[
\chi^2 = \sum_{i=1}^{16} \frac{(D_i - S_i)^2}{S_i},
\]

Here, we have assumed that the error in each bin is dominated by Poisson error. Our Markov chain is run for 4000 iterations, with a “burn in” of 200 steps that are excluded from subsequent statistical analysis. A new set of parameters are accepted if \(\chi^2_n < \chi^2_{n-1}\) (also known as Gibbs sampling), where the index refers to the Markov step. If \(\chi^2_n > \chi^2_{n-1}\), the algorithm accepts
the new parameter set with probability $e^{-\frac{(x_0-x_n)^2}{2}}$. We adopt the medians of the accepted steps as the best-fit set, and we calculate both upper and lower standard errors using the 16th and 84th percentile values. The error of the $2.8 < R_p < 4$ bin is calculated from the standard deviation of $\mu_{\chi^2}$, i.e., the relative occurrence of the $2.8-4 R_\oplus$ bin with respect to the $2-2.8 R_\oplus$ bin.

2.5. Iterative Simulation (IS)

We use the method of sequential Monte Carlo with bootstrap filter described in Cappé et al. (2007) to infer the intrinsic planet radius distribution without resorting to binning. In this technique we first generate a trial population of planets by simulating detections (Section 2.2). The radii of these simulated detections are then replaced by actual KOIs, and the process repeats until the simulated detections converge on the observations. Convergence is established when the radius distribution of our simulated detected planets (see dotted line in Figure 2) matches the observed 430KOI distribution. The radius distribution of the trial population then reflects that of the intrinsic population of planets. With a sufficiently large trial population, the resolution of the radius distribution is limited only by the amount of information in the observations (i.e., KOIs). In principle the periods can also be replaced by actual KOI values, but we instead choose to fix the distribution of period values using Equation (2) and $\alpha$, determined from our MCMC analysis. Since we find an excellent $\alpha$ fit to the period distribution (Section 3.2) over $20 < P < 200$, applying this distribution instead reduces additional noise in our simulation.

Our IS simulates $10^6$ transiting planet–star pairs, drawing stars from the selected catalog with replacement (Section 2.2). Eccentricities and arguments of periastron of trial planets are drawn from Rayleigh and uniform distributions, respectively, and periods are drawn from a power law with the index of the best-fit MCMC model (Section 2.4), $\alpha = -0.04$. Planet radii are initially drawn from a uniform distribution over $0.5-6 R_\oplus$. Detections are simulated as described in Section 2.2. We randomly replace the radii of all simulated detected planets with values drawn from the “master” radius distribution (Section 2.3), after correcting for false positives using the rates in Table 1 of Fressin et al. (2013). For a discussion of our false positive treatment, see Section 5.1. We then redraw new values for all the other planet parameters besides radius and period, and reschedule the planets among the stars. We repeat this process until acceptable convergence is achieved, usually within 100 iterations. At this point, the trial population is used to calculate the intrinsic radius distribution.

Errors are calculated by constructing 50 bootstrapped samples of the detected planet catalog. The size of each sample is a random Poisson deviate with expectation equal to the size of the actual sample. The bootstrapped samples are drawn with replacement from the actual KOI sample. Planets are randomly removed according to the false positive probabilities of Fressin et al. (2013) and then intrinsic radius distributions are calculated as in Section 2.3. For our bootstrapped samples we make the false positive correction before constructing the much larger intrinsic distribution to capture the contribution of false positives to the “noisiness” of each bootstrapped sample. We analyze each sample using the IS technique and compute standard deviations of the ensemble of bootstrapped planet populations to represent 1σ uncertainties.

Figure 2 shows the results of an artificial test case of the reconstruction of a planet radius distribution using the IS technique. The intrinsic distribution (dashed line) is the sum of a Gaussian plus a rising slope. The dotted line is the distribution of 364 simulated observations, which is similar in scale to our 430KOI sample. We apply the IS technique on these simulated observations to recover the actual distribution: the result is plotted as the solid line, with error bars determined from 25 bootstrapped runs. The ability of IS to reconstruct an intrinsic distribution is limited by the information available in any region of a distribution, i.e., it will fail where the number of planets or rate of detection is too low. In Figure 2, errors or large uncertainties appear in the reconstructed at $R_p \sim 1 R_\oplus$ where the detection efficiency is low. Also, in this simple demonstration we ignore the effect of planet radius errors. Adding planet errors tends to broaden and smooth features.

2.6. Calculation of Occurrence

After obtaining best-fit distributions from our MCMC and IS analyses, we calculate the rate of planet occurrence as a function of $P$ and $R_p$. We generate $10^6$ mock star–planet pairs and the corresponding simulated detections (Section 2.2). These planets are binned in a logarithmic grid of $P$ (index $i$) and $R_p$ (index $j$). The occurrence $f(i, j)$ in the bin $(i, j)$ is then

$$f(i, j) = \frac{K_{ij} N_{\ast,S}}{S},$$

where $K_{ij}$ is the false positive-corrected (Table 1, Fressin et al. 2013) number of KOIs falling into bin $(i, j)$, $N_{\ast} (= 76,711)$ is the total number of Kepler target stars in the Solar76k sample, $N_{\ast,S} (= 10^6)$ is the total number of mock pairs, and $S$ is the number of simulated detections. The fraction of mock pairs that should be detected is $S/N_{\ast,S}$ and is the product of the geometric factor $R_\ast/a$ as well as the detection completeness in bin $(i, j)$, hereafter known as $C(i, j)$ (see Section 3.1). We then sum over all bins to obtain the total occurrence, $f$.

To demonstrate the importance of accounting for radius errors (Section 2.3) when calculating occurrence, we also conduct a separate analysis that excludes radius errors. Observed planets are binned as above and we calculate the occurrence of each
Section 3.1. Completeness

For a transit survey, completeness is the fraction of transiting planets of a given P and Rp that are actually detected, i.e., not including the geometric transit probability. Accurately capturing the dependence of completeness on Rp and P is crucial to a robust determination of planet occurrence.

We emphasize that survey completeness depends not only on the properties of the planet host stars, but also on the stellar and noise properties of the entire catalog. We calculate the completeness C(i, j) of bin (i, j) by inserting 2 × 104 planets randomly distributed around the Solar76k sample of stars. The fraction of “detected” planets, modulo the transit probability factor, yields the completeness in this bin. The results are displayed in Figure 3.

Figure 3 shows that Kepler completeness is nearly 100% for planets larger than Neptune (3.8 R⊕) and for nearly the full range of periods shown here. This falls rapidly beyond P ∼ 500 days (not shown) since some systems no longer have the required three transits during the four-year Kepler mission. The completeness drops rapidly with decreasing radius where R < 2 R⊕: Earth-sized planets are readily detected by Kepler only if they have orbital periods of a few days, and beyond P = 200 days, even the completeness of 2 R⊕ planets falls below 50%. For these reasons we restrict our analysis to P < 200 days. As true of any completeness study, we also note that our completeness calculations are based on the S/N criterion as stated. For example, if the Kepler pipeline was uniformly missing 25% of all transiting planets (regardless of the S/N), our completeness calculations would not account for these missing planets.

3. THE PRIMORDIAL POPULATION OF Kepler PLANETS

3.1. Completeness

The numbers within each grid cell indicate the completeness percentage, and each grid has been color coded from low (blue) to high (white) completeness.

\[
f(P_i, R_i) = \frac{1}{N_s} \sum_k \left( \frac{a_k}{R_{s,k}} \right) C(i, j), \tag{11}\]

where C(i, j) is the average completeness of the bin (see Section 3.1), n_p(i, j) is the number of planets in bin (i, j), N_s (≈ 76,711) is the total number of stars in the sample, and \(a_k/R_{s,k}\) is the geometric correction factor for planet k.

3.2. Period Distribution

Our MCMC study yields a best-fit distribution for the 430KOI sample of \(\alpha = -0.04 \pm 0.09\) (see Equation (2)), with a reduced chi-squared \(\chi^2\) of 1.07. Our value of \(\alpha\) is consistent with zero (a flat logarithmic distribution) within errors, confirming previous determinations (e.g., Youdin 2011; Howard et al. 2012; Petigura et al. 2013; Fressin et al. 2013).

Figure 4 compares this best-fit period distribution with the observed sample, using the four period bins from Section 2.4 as well as an additional two bins to include to include planets inward of 20 days, i.e., the 1052KOI sample (Section 2.1). Inside of P = 20 days, Kepler planets deviate from a simple power-law distribution (also see Youdin 2011; Howard et al. 2012).

3.3. Radius Distribution

Figure 5 displays our best-fit radius distributions for both the MCMC (solid, black) and IS (dotted, red) techniques, where we have binned the IS result for ease of comparison. Both the IS and MCMC distributions peak at 2–2.8 R⊕ and decrease toward smaller radii. We have also plotted two additional distributions in Figure 5, a “No Error” case (dashed, green) constructed from Equation (11) and a “25% Larger” IS case (dash-dotted, blue) where it is assumed that both planet and stellar radii are 25% larger than their catalog values.

As we discuss in Section 2.3, errors in the radius of candidate Kepler planets, primarily due to uncertainties in stellar radius, are large and detection bias against small planets means that a planet’s cataloged radius is likely an underestimate. Comparing our IS and MCMC results (which include radius errors) with our “No Error” case (which does not include radius errors) we see that the latter exhibits a significant excess of 1–1.4 R⊕ planets. This is as expected. Correcting for radius error in a Bayesian way (Section 2.3) tends to promote small planets to larger size bins, and depopulates the smallest radius bin. However, planet occurrence of the two largest bins does not increase significantly since the survey completeness is substantially higher in these bins compared to the 1–1.4 R⊕ bin. We conclude.
that not accounting for this detection bias on radius leads to an erroneously high (by a factor of ∼2) value of occurrence for the 1–1.4 $R_\oplus$ bin. If we assume the extreme scenario that the true radii of all stars (and therefore their planets) are 25% larger than the KOI values (“25% Larger” case in Figure 5), as is shown to be the case for at least a subset of the Kepler stars (Section 5.1), we observe that the 2–2.8 $R_\oplus$ peak is now shifted to 2.8–4 $R_\oplus$. However, we do not see a significant change in the bin 1–1.4 $R_\oplus$ because the depopulation of this region (due to increased planet radii) is roughly balanced by a decrease in completeness as the stars have also become larger. See Section 5.1 for a more detailed discussion.

Our treatment for the radius error is far from perfect. Some genuinely small planets may, under our procedure, be wrongly inferred to have larger radii. A better treatment will require improved errors in stellar/planet radius (see, e.g., Section 5.3). There is a small and statistically insignificant discrepancy between the MCMC and IS results at the smallest bin ($1 < R_p < 1.4 \; R_\oplus$). Since the two methods use identical input catalogs (Section 2.1) and detection algorithms (Section 2.2), the difference could be due to the intrinsic binning in the MCMC method.

We test the effect of different bin sizes on our MCMC results by varying the bin sizes in both radius and period space. Using the logarithmic binning scheme to describe the width, $w$, of each bin:

$$w(n) = 10^K_n,$$

where $n$ is the bin number and $K$ is a constant, we investigate the effect of varying $K$ (and thus bin size) on occurrence. When varying period bin size, $K_p$, and keeping the radius bin size, $K_r$, constant we find no effect on occurrence, even when $K_p$ is varied by a factor of two. However, when $K_p$ is kept constant and $K_r$ is varied, we do find an affect on occurrence rates shown in Figure 6 (where our IS result from Figure 5 is plotted as a dotted black line). As $K$ is increased (larger bins), we see that the occurrence of $2 < R_p < 2.8 \; R_\oplus$ planets increases while the occurrence of $2.8 < R_p < 4 \; R_\oplus$ planets decreases, becoming significantly different from our IS result for extreme cases. In contrast, there is no statistically significant difference in the smallest bin. We find that the error bars increase with more extreme bin sizes, indicating that the MCMC algorithm has a harder time converging. The default bin sizes for the MCMC result in Figure 5 are $K_p = 0.301$ and $K_r = 0.150515$.

Lastly, we display the IS radius distribution for a smaller logarithmic bin size in Figure 7. As explained in Section 2.5, since the IS technique requires no binning, the resolution of the result is limited only by the data and its errors. This improved resolution can reveal finer detail about the intrinsic radius distribution. We observe a slight excess of planets in the now smallest bin $(1–1.15 \; R_\oplus)$, over that in larger bins, though
improved statistics are required to confirm this upward turn. We expect that with its independence on binning, the IS technique will become central to future analysis.

3.4. Total Occurrence of Small Planets

In Table 1 we report our estimates for the total planet occurrence within 20 < P < 200 days and 1 < R_p < 4 R_⊕, for the four curves in Figure 5. There is excellent agreement between the MCMC and the IS results. Even cases with different assumption about the radius error yield statistically consistent occurrence rates.

The total occurrence rate we calculate here is defined to be the average number of planets per star (Youdin 2011; Fressin et al. 2013; Petigura et al. 2013). Such a definition ignores the complication that many of the Kepler systems are multiple systems (e.g., Lissauer et al. 2011). A quantity perhaps more relevant for studies of planet formation is the occurrence rate of planetary systems, or the average number of planetary systems a star has. However, this requires knowledge of the system architecture, a task not yet attempted.

Lastly, if we include small planets on orbits interior to 20 days, the total occurrence rate is raised to ~66%. It will rise by another ~15% if we include planets larger than 4R_⊕.

3.4.1. Eta-Earth

We estimate \( \eta_\oplus \), the occurrence of Earth-like planets in the “habitable zone” of solar-type stars. By Earth-like, we are referring to planets between 1–2 R_⊕. We adopt the inner and outer boundaries of the habitable zone to be those calculated by 1D, cloud-free, climate models (Kasting et al. 1993; Kopparapu et al. 2013). For a star with a Solar spectrum, these boundaries lie at 0.99 and 1.70 AU respectively (or orbital periods of 350 and 810 days); for other stellar spectral types, the boundaries are as tabulated in Kopparapu et al. (2013).

Such a habitable zone, however, lies outside the 200 day limit of our study. At these very long periods, the low detection efficiency of Kepler engenders inaccuracies in estimating \( \eta_\oplus \). So instead, we have opted to calculate \( \eta_\oplus \) by extrapolation, according to

\[
\eta_\oplus = \frac{f}{N_\star} \sum h_i, \tag{13}
\]

where \( N_\star = 76,711 \) is the number of stars in the Solar76k sample and \( h_i \) is the relative occurrence of planets (per star) within the habitable zone to some reference zone having absolute occurrence \( f \). We choose this reference zone to be our standard 20 < P < 200 days, 1 < R_p < 4 R_⊕ bound, with \( f = 0.46 \pm 0.03 \) (IS value). We choose this period because the stars change positions on the detector space our completeness values are slightly and systematically higher than those of PHM13, and the discrepancy increases with increasing \( P \) and decreasing \( R_p \). This is to be expected because PHM13 determine detection efficiency using actual light curves rather than representations of noise a la CDPP values. At \( P > 200 \) days our values of \( C \) become significantly higher than PHM13 for nearly all values of \( R_p \). This

\[
\eta_\oplus = 6.4^{+3.4}_{-1.1}\%. \tag{14}
\]

The error is calculated from error propagation of the IS radius distribution, occurrence of our reference zone, habitable zone limits (based on \( R_\star, M_\star, \) and \( T_\star \)) and \( \alpha \). This value is consistent within errors with the analysis done by PHM13 for the same Kopparapu et al. (2013) limits, 8.6%. The reader is reminded that our calculation of \( \eta_\oplus \) is an extrapolation, and depends crucially on the assumptions made.

4. COMPARISON WITH PHM13

We now compare our work to PHM13—an analysis that is similar to ours in terms of scope but which obtains their results of the Kepler data using the TERRA pipeline (Petigura et al. 2013). The TERRA pipeline is an analysis tool independent of the Kepler project pipeline and its products on which our work relies.

We first compare our estimates of detection completeness \( C \) with that of PHM13. For this completeness comparison, we re-compute \( C \) using the Best42k stars from PHM13 and compare these results to the values in Figure S11 from PHM13. We calculate the fractional difference \( (2(P - T(P))/T(P) \), where \( T_\star = This Work and \( P = PHM13 \)) and display as percentages in Figure 8. With the exception of a single cell all values of \( C \) are within \( \pm 1\% \) of those reported by PHM13 in their Figure S11, expressed as a fractional difference (i.e., in percentages). Our values are generated using the “Best42k” catalog from PHM13 and our detection criteria. Bins where both our completeness and those of PHM13 are zero have been blacked out.

\[
\eta_\oplus = 6.4^{+3.4}_{-1.1}\%. \tag{14}
\]

The error is calculated from error propagation of the IS radius distribution, occurrence of our reference zone, habitable zone limits (based on \( R_\star, M_\star, \) and \( T_\star \)) and \( \alpha \). This value is consistent within errors with the analysis done by PHM13 for the same Kopparapu et al. (2013) limits, 8.6%. The reader is reminded that our calculation of \( \eta_\oplus \) is an extrapolation, and depends crucially on the assumptions made.

4. COMPARISON WITH PHM13

We now compare our work to PHM13—an analysis that is similar to ours in terms of scope but which obtains their results of the Kepler data using the TERRA pipeline (Petigura et al. 2013). The TERRA pipeline is an analysis tool independent of the Kepler project pipeline and its products on which our work relies.

We first compare our estimates of detection completeness \( C \) with that of PHM13. For this completeness comparison, we re-compute \( C \) using the Best42k stars from PHM13 and compare these results to the values in Figure S11 from PHM13. We calculate the fractional difference \( (2(P - T(P))/T(P) \), where \( T_\star = This Work and \( P = PHM13 \)) and display as percentages in Figure 8. With the exception of a single cell all values of \( C \) are within \( \pm 1\% \) of those reported by PHM13 in their Figure S11, expressed as a fractional difference (i.e., in percentages). Our values are generated using the “Best42k” catalog from PHM13 and our detection criteria. Bins where both our completeness and those of PHM13 are zero have been blacked out.

\[
\eta_\oplus = 6.4^{+3.4}_{-1.1}\%. \tag{14}
\]

The error is calculated from error propagation of the IS radius distribution, occurrence of our reference zone, habitable zone limits (based on \( R_\star, M_\star, \) and \( T_\star \)) and \( \alpha \). This value is consistent within errors with the analysis done by PHM13 for the same Kopparapu et al. (2013) limits, 8.6%. The reader is reminded that our calculation of \( \eta_\oplus \) is an extrapolation, and depends crucially on the assumptions made.

4. COMPARISON WITH PHM13

We now compare our work to PHM13—an analysis that is similar to ours in terms of scope but which obtains their results of the Kepler data using the TERRA pipeline (Petigura et al. 2013). The TERRA pipeline is an analysis tool independent of the Kepler project pipeline and its products on which our work relies.

We first compare our estimates of detection completeness \( C \) with that of PHM13. For this completeness comparison, we re-compute \( C \) using the Best42k stars from PHM13 and compare these results to the values in Figure S11 from PHM13. We calculate the fractional difference \( (2(P - T(P))/T(P) \), where \( T_\star = This Work and \( P = PHM13 \)) and display as percentages in Figure 8. With the exception of a single cell all values of \( C \) are within \( \pm 1\% \) of those reported by PHM13 in their Figure S11, expressed as a fractional difference (i.e., in percentages). Our values are generated using the “Best42k” catalog from PHM13 and our detection criteria. Bins where both our completeness and those of PHM13 are zero have been blacked out.

\[
\eta_\oplus = 6.4^{+3.4}_{-1.1}\%. \tag{14}
\]

The error is calculated from error propagation of the IS radius distribution, occurrence of our reference zone, habitable zone limits (based on \( R_\star, M_\star, \) and \( T_\star \)) and \( \alpha \). This value is consistent within errors with the analysis done by PHM13 for the same Kopparapu et al. (2013) limits, 8.6%. The reader is reminded that our calculation of \( \eta_\oplus \) is an extrapolation, and depends crucially on the assumptions made.

4. COMPARISON WITH PHM13

We now compare our work to PHM13—an analysis that is similar to ours in terms of scope but which obtains their results of the Kepler data using the TERRA pipeline (Petigura et al. 2013). The TERRA pipeline is an analysis tool independent of the Kepler project pipeline and its products on which our work relies.

We first compare our estimates of detection completeness \( C \) with that of PHM13. For this completeness comparison, we re-compute \( C \) using the Best42k stars from PHM13 and compare these results to the values in Figure S11 from PHM13. We calculate the fractional difference \( (2(P - T(P))/T(P) \), where \( T_\star = This Work and \( P = PHM13 \)) and display as percentages in Figure 8. With the exception of a single cell all values of \( C \) are within \( \pm 1\% \) of those reported by PHM13 in their Figure S11, expressed as a fractional difference (i.e., in percentages). Our values are generated using the “Best42k” catalog from PHM13 and our detection criteria. Bins where both our completeness and those of PHM13 are zero have been blacked out.

\[
\eta_\oplus = 6.4^{+3.4}_{-1.1}\%. \tag{14}
\]

The error is calculated from error propagation of the IS radius distribution, occurrence of our reference zone, habitable zone limits (based on \( R_\star, M_\star, \) and \( T_\star \)) and \( \alpha \). This value is consistent within errors with the analysis done by PHM13 for the same Kopparapu et al. (2013) limits, 8.6%. The reader is reminded that our calculation of \( \eta_\oplus \) is an extrapolation, and depends crucially on the assumptions made.
discrepancy motivates our restriction to $P < 200$ days. One possible explanation for this difference is that detection of signals by phase folding in the *Kepler* detection pipeline becomes inefficient at long periods.

We note that PHM13 calculates completeness based on a finite number ($4 \times 10^4$) of systems of which very few detections are in the Earth-sized bins, leading to large counting (Poisson) error in completeness values. Since we use a different approach, simulating large numbers ($2 \times 10^6$ total, $2 \times 10^4$ per bin) of test planets and giving high drawing probability to Earth-sized planets, we are able to simulate a much larger number of detections for each bin and thus have a more precise (although not necessarily more accurate) value of completeness.

We next compute the impact of these differences in completeness on occurrence over $P = 5\text{--}100$ days, shown in Figure 9 (note that the radius errors of Section 2.3 have been omitted here). We first calculated occurrence using the “603PHM” data set (Section 2.1), Equation (11) along with our own detection criteria (Section 2.2) and completeness values (calculated from the Best42K sample), shown as the solid black curve in Figure 9. We then recalculated planet occurrence using the 603PHM data set, Equation (11) along with our own detection criteria but substituting in the completeness values from Figure S11 of PHM13 for ours, shown as the dashed green line in Figure 9. Lastly, the direct results of PHM13 are shown as a dotted, red line. The differences between all curves in Figure 9 are small and, within errors, agree with each other. The residual differences in completeness seen in Figure 8 do not appear to play a significant role in the comparative occurrence of our work and that of PHM13 but could be responsible for some of the minor (and statistically insignificant) differences that we find. We conclude that simple detection criteria and noise model can be used in planet occurrence studies to achieve accurate and precise results.

Comparing PHM13’s intrinsic radius distribution (red curve in Figure 9) to our own (black curve in Figure 5), we notice a difference in the occurrence of large ($2 < R_p < 4$) planets. Specifically, PHM13 calculated an occurrence of $18.5 \pm 1.5\%$ and $6 \pm 0.5\%$ for $2 < R_p < 2.8$ and $2.8 < R_p < 4$ planets, respectively, while we find an occurrence of $13.0 \pm 1.1\%$ and $10.5 \pm 1.0\%$ for the same bins. This results in a statistical difference of about 2% and 3% between works for the $2 < R_p < 2.8$ and $2.8 < R_p < 4$ bins, respectively. We see a few possible reasons for this. Table 2 displays the major differences in raw samples between our work and PHM13, displaying the raw occurrence of large ($2 < R_p < 4$) planets for their nominal period ranges, $P > 20$ days and $P < 20$ days. The bottom two rows of “This Work” are empty in Table 2 since all of the planets in our sample are $P > 20$, and thus all the relevant information is present in the first row.

To summarize Table 2, of the 394 KOIs in our raw planet sample between $1 < R_p < 4$, 166 (42%) and 120 (30%) are between $2 < R_p < 2.8$ and $2.8 < R_p < 4$, respectively, while for PHM13, of the 495 KOIs between $1 < R_p < 4$ only 191 (39%) only fall into the same limits, respectively. This is a significantly lower fraction, and it appears that our sample and PHM13’s sample are markedly different. There are only a couple options available to explain this difference—either a disproportional number of large ($2 < R_p < 4$) planets in our sample are false positives or the two data sets are not subsamples from the overall same population.

If the two samples are from different populations, this difference we claim is due to the photoevaporation (Owen & Wu 2013) of PHM13’s close-in planets, which acts to convert large ($2 < R_p < 4$) planets into smaller ones. Splitting the PHM3 data set into $P > 20$ day and $P < 20$ day subsets adds support to this theory, since as we move from small to large periods raw occurrence drops proportionally by (44%--35%)/44% ~ 20% and (22%--14%)/22% ~ 40% for $2 < R_p < 2.8$ and $2.8 < R_p < 4$ planets, respectively.

In explaining the difference between the occurrence of large planets between this work and PHM13, one must also consider the improved treatment of radius errors and updated Huber et al. (2014) stellar parameters used in this work. Both tend to increase planet size, pushing smaller planets into larger bins. Comparing the “No error” case to the “IS” case in Figure 5, we can see that the former effect increases the occurrence of $2 < R_p < 2.8$ and $2.8 < R_p < 4$ planets by about 2% and 1.5%, respectively. Thus, it appears that the discrepancy of our results with the PHM13 in the $2 < R_p < 2.8$ bin can be explained by our incorporation of radius errors. However this does not account for the total discrepancy in the $2.8 < R_p < 4$ bin, leaving about 1.5% of discrepancy between our work and PHM13.

A final source of discrepancy in the $2.8 < R_p < 4$ bin between works may come from the different stellar populations

### Table 2

| Range               | This Work | PHM13 |
|---------------------|-----------|-------|
| Nominal $P$ range:  |           |       |
| $1 < R_p < 4$       | 394       | 495   |
| $2 < R_p < 2.8$     | 166 (42%) | 191 (39%) |
| $2.8 < R_p < 4$     | 120 (30%) | 87 (18%) |
| $P > 20$ days:      |           |       |
| $1 < R_p < 4$       | ...       | 201   |
| $2 < R_p < 2.8$     | ...       | 89 (44%) |
| $2.8 < R_p < 4$     | ...       | 45 (22%) |
| $P < 20$ days:      |           |       |
| $1 < R_p < 4$       | ...       | 294   |
| $2 < R_p < 2.8$     | ...       | 102 (35%) |
| $2.8 < R_p < 4$     | ...       | 42 (14%) |
used between our work and PHM13. Mulders et al. (2014) found that the occurrence of planets is correlated with stellar type, and an analysis of the Best42k sample shows that ~30% of the stars used in the PHM13 analysis are outside the 0.8 < R_p/R_⊕ < 1.2 range. However, it should be noted that Mulders et al. (2014) does not predict a significant change in occurrence across stellar types for 2.8 < R_p < 4 R_⊕ planets.

To conclude this comparison, PHM13 reported an occurrence of 37 ± 3.4% for planets with 25 < P < 200 days and 1 < R_p < 4 R_⊕. Including planets from 20 < P < 25 days raises this value to 42 ± 3.6%. So the overall occurrence rates are consistent among studies that are based on different detection criteria and different model assumptions.

5. DISCUSSION

5.1. Sensitivities and Systematics

To investigate sensitivities to some of the assumptions and parameters in our analysis, we repeat our IS simulations varying σ_e and α. First, we varied the Rayleigh distribution of orbital eccentricities between 0.1 and 0.3, and second varied the value of α between −0.15 and 0.15. In all cases we found no significant difference in the radius distributions. We also investigate our separability assumption (Equation (1)) by splitting our 430KOI data set into two equal sized subsets corresponding to 20 < P < 50 and 50 < P < 200 day planets, and performing an MCMC analysis on each. The resulting distributions are consistent with each other as well as with our main IS and MCMC results, indicating that there is no significant correlation between planet radius and period in our sample, and that Equation (1) is a reasonable assumption.

In this analysis, we address the fact that many Kepler planets have large errors in radius, driven primarily by uncertainties in the radii of their host stars. But we have not addressed the issue of systematic errors in stellar radii. For example, stellar effective temperatures based on photometry from KIC (Brown et al. 2011) are systematically ~200 K hotter than more reliable estimates based on the infrared flux method (Pinsonneault et al. 2012) or spectroscopy (Gaidos 2013). The combination of uncertainties in stellar parameters and Malmquist bias in the magnitude-limited Kepler target catalog means that the sample is biased toward the most luminous, hottest, and largest stars (Gaidos & Mann 2013). There is increasing evidence that many Kepler target stars, including planet hosts, are subgiants (e.g., Verner et al. 2011; Everett et al. 2013; Bastien et al. 2014). For fixed values of R_p/R_*, systematically larger stellar radii means the planets are also systematically larger and that the geometric transit probability is higher than presumed (transit probability depends inversely on stellar density and hotter, more evolved stars are less dense). The detection completeness of small planets is also smaller than presumed and thus, for fixed number of detections, the occurrence is higher.

We have explored the possible impact of these effects by assuming that all stellar radii in our Solar76k catalog, as well as all the planet radii in the corresponding 450KOI catalog, are 25% larger than their nominal values (dash-dotted blue curve; Figure 5). The distribution differs markedly from our IS and MCMC distributions. The peak in the distribution at 2–2.8 R_⊕ has shifted toward larger radii. Surprisingly, the occurrence of planets with R_p = 1–1.4 R_⊕ has not changed. This is understood. As stellar radii increase, completeness of small planets decrease leading to an increase in planet occurrence. However, the number of 1–1.4 R_⊕ planets in our sample also decreases (from 25 to 8 after a 25% radius increase), reducing the raw planet occurrence in this radius bin. It appears that these competing effects roughly cancel, resulting in no significant change in planet occurrence of the smallest radius bin.

We now comment on our treatment of false positives in this analysis. Our work makes use of the Fressin et al. (2013) false positive rates based on the Q1–Q6 Kepler data, while other works (e.g., PHM13) use custom methods to detect false positives. However since we use the latest disposition of the Kepler catalog in this analysis, the occurrence of false positives in our Solar76k sample could be significantly different. We calculate the ratio of false positives ("FP") vetted by the Kepler science team to planet candidates ("CAN") for the Q1–12, Q1–16, and cumulative Kepler catalogs according to FP/(FP+CAN). In addition, we organize these false positive ratios by radius, using the same radius bins as our analysis. These false positive ratios are shown in Table 3, along with the Fressin et al. (2013) values for reference. It should be noted that the fraction of planets yet to be dispositioned (calculated according to DISP/(DISP+CAND+FP), where DISP is the number of planets yet to be dispositioned) in the Q1–12 and Q1–16 data sets are quite high (over 50 %), and may reflect the significant difference between their false positive rates and the cumulative KOI data set rates.

We use the false positive fractions in Table 3 to estimate the uncertainty of using Fressin et al. (2013) false positive rates in our analysis. For Q1–12, Q1–16, and the cumulative list, we carry out separate MCMC analyses substituting in the false positive rates from Table 3. Only the cumulative list remains consistent with our main IS and MCMC results, with overall occurrence rates of the Q1–12, Q1–16, and cumulative data sets being 0.30 ± 3%, 0.32 ± 3%, and 0.38 ± 3%, respectively, i.e., much lower occurrence values. Taking the standard deviation of each radius bin (using our main IS result plus the three new results derived from the Q1–12, Q1–16, and cumulative positive rates), we estimate the uncertainty in our false positive rates on the absolute occurrence of planets for 1 < R_p < 1.4, 1.4 < R_p < 2, 2 < R_p < 2.8, and 2.8 < R_p < 4 R_⊕ to be 2.1%, 1.6%, 1.8%, and 2.4%, respectively.

5.2. Astrophysical and Astrobiological Implications

The period distribution of Kepler small planets contains two distinct parts. The first is a rise from ~5–10 days (e.g., Youdin 2011; Howard et al. 2012), the second is a logarithmically flat distribution extending from ~10 days out to at least 200 days (Figure 4 here; Petigura et al. 2013; Fressin et al. 2013). The origin of both features are unclear. However, we speculate on one origin for the logarithmically flat feature. Imagine a set of planetary systems comprised of closely packed, equal-mass planets. Dynamical stability requires that neighboring planets be spaced apart by more than a few Hill radii (Chambers et al. 1996; Smith & Lissauer 2009). Since the Hill radius scales linearly with orbital semimajor axis, this means the separation between neighboring planets grows linearly with their orbital span. This would then translate into a period distribution that is
flat in logarithmic period. In other words, it is possible that most or all of our planets are actually in multiple systems, and that the flat feature is a result of the stability requirement. Although Fang & Margot (2012) have quoted that 75–80% of planetary systems have one or two planets with orbital periods less than 200 days (suggesting that most systems are not close to the Hill stability limit), this result is based on the occurrence of observed systems. Most Earth-sized planets (or smaller) around Sun-like stars are undetectable by Kepler, and it is possible that the multiplicity of such systems is much higher than we currently believe due to these unseen planets.

Alternatively, the flat feature can arise from the primordial mass distribution in the disk. Assuming that all planets have comparable masses, are in multiple systems, and are formed where they are found today, a logarithmically flat spacing would suggest that the disk surface density $\Sigma$ scales with the orbital separation $a$ as

$$\Sigma \propto a^{-2}. \quad (15)$$

This is not vastly different from the theoretical minimum-mass solar nebula (MMSN) profile: $\Sigma \propto a^{-3/2}$ (Hayashi 1981; Weidenschilling 1977), a useful benchmark to study protoplanetary disks.

The radius distribution is equally intriguing. The radius of an Earth- to Neptune-sized planet mostly reflects the expanse of its hydrogen envelope (Wolfgang & Lopez 2014). By focusing on planets outward of 20 days, we discard candidates that may have had their atmospheres eroded by stellar irradiation Owen & Wu (2013). The distribution shown in Figure 5 is therefore likely “primordial.” Compared to planets inward of 10 days that have radii $\lesssim 1.5 \, R_\oplus$, this “primordial” population appears to prefer a size of $\sim 2.5 \, R_\oplus$. Such a size corresponds to a fractional mass in the hydrogen envelope of $\sim 1\%$ (assuming a rocky core roughly in the 10 $M_\oplus$ range, see, e.g., Wu & Lithwick 2013). What is the reason behind this preference for 1%? A planet embedded in a proto-planetary disk can accrete a hydrostatic atmosphere. Rafikov (2006) calculated that this atmosphere has a mass of a few $M_\oplus$ for a 10 $M_\odot$ planet at 0.1 AU in a MMSN disk. This lies much above the 1% value but it depends on disk parameters and its evolution history. In future works, the observed radius distribution should be used to decipher formation history.

Moreover, the gradual decline toward smaller sizes in logarithmic space has implication for the formation of bare-core planets, the norm in the inner solar system. Our terrestrial planets are thought to have formed in a gas-free environment by conglomeration of solid materials. The relative shortage of bare-core planets may suggest that the observed Kepler planets may have followed different formation path than that of the terrestrial planets.

Lastly, we turn to the issue of $\eta_B$. We calculate $\eta_B$ more out of respect for tradition than with any conviction that there is additional accuracy to be assigned to our calculation. The limits of the habitable zone depend on important assumptions regarding the climate state of Earth-like planets (Kopparapu et al. 2013), mass (Kopparapu et al. 2014), and the composition of the atmosphere (Pierrehumbert & Gaidos 2011). Nevertheless, the search for life elsewhere can take heart in the fact that multiple investigations point to an occurrence of Earth-size planets in habitable zones of $O(0.1)$ or more. Indeed, studies of M dwarfs suggest that $\eta_B \sim 0.5$ (Bonfils et al. 2013; Kopparapu 2013; Gaidos 2013). M dwarfs comprise about 70% of all stars and hence weigh heavily in the census for Earth-like planets.

### 5.3. Improvements in Occurrence will Happen

As seen in Table 4, the errors associated with most Kepler planets are dominated by the uncertainty in the parameters of their host stars. Thus, in order to improve planet occurrence calculations for the future we must first understand Kepler stars better. The Gaia (Global Astrometric Interferometer for Astrophysics) mission, launched in 2013 December, will measure the parallaxes of 1 billion stars in the local group with accuracies approaching 10 μas, as well as obtain multi-band photometry measurements (de Bruijne 2012). Liu et al. (2012) estimate that for stars in the KIC, Gaia will be able to estimate $T_{\text{eff}}$ to 1%, log g to within 0.1–0.2 dex, and [Fe/H] to within 0.1–0.2 dex. The combinations of these data should dramatically improve our knowledge of the properties of Kepler target stars and hence reconstructions of the Kepler planet population.

Other advances include improved maps of interstellar reddening in the Kepler field based on the colors of oscillating red giants with established properties, as well as WISE infrared photometry (Huber et al. 2014). The advent of multiplexed, multi-object spectrographs capable of simultaneously measuring thousands of stars (Hill et al. 2010) should, combined with Gaia parallaxes, allow stellar parameter estimation with unprecedented scale and precision. In addition, measurement of photometric noise due to stellar granulation (“flicker”) is a promising technique for estimating the log g and hence radius of bright Kepler stars to within 0.1–0.2 dex (Bastien et al. 2014), although its calibration and applicability to fainter Kepler stars—the majority of targets, with lower photometric precision—remains to be seen.

### 6. CONCLUSIONS

In this work we have developed a population simulator to extract the underlying period and radius distributions of Earth- to Neptune-sized planets detected by Kepler. We focus on a “primordial” population of planets outside 20 days to

---

**Table 4**

The 430KOI Planet Sample

| KOI       | Period | $R_p/R_\oplus$ | $R_\oplus$ | $\sigma_{R_\oplus}$ | $\sigma_{R_p}$ | $R_p$ | $\sigma_{R_p}$ | $S/N$ |
|-----------|--------|----------------|------------|----------------------|----------------|-------|----------------|-------|
| K00435.05 | 62.3   | 0.02           | 0.83       | 0.34                 | 0.06           | 2.47  | 0.83           | 1.32  |
| K02289.02 | 20.0   | 0.01           | 1.13       | 0.70                 | 0.13           | 1.45  | 1.62           | 1.85  |
| K00880.01 | 26.4   | 0.04           | 0.92       | 0.38                 | 0.08           | 4.09  | 0.37           | 1.68  |
| K00880.02 | 51.5   | 0.05           | 0.92       | 0.38                 | 0.08           | 5.47  | 0.49           | 2.25  |
| K04150.01 | 31.3   | 0.01           | 1.17       | 0.55                 | 0.14           | 1.94  | 3.05           | 3.18  |

Note. The radii of these planets have been updated with Huber et al. (2014) parameters where applicable. (This table is available in its entirety in machine-readable form.)
exclude the impact of, e.g., photoevaporation. We find that the adoption of a simple model of photometric noise and transit signal detection allow us to accurately estimate the survey completeness of Kepler. We have accounted for radius errors in our analysis, and have found that doing so is important for reconstructing the intrinsic radius distribution. We apply the IS technique to reconstruct the planet distribution with radius. This does not require binning and allows radius errors to be readily accounted for. Last, we are the first to use the updated Huber et al. 2014 parameters along with all 16 quarters of Kepler data, representing the most up to date analysis. The main results are as follows.

1. The distribution of planets with $20 < P < 200$ days is roughly uniform with logarithmic period (power-law index $\alpha = -0.04 \pm 0.09$).

2. The (likely primordial) radius distribution for Kepler planets with $20 < P < 200$ days peaks in the radius bin $2-2.8 R_\oplus$.

3. The overall occurrence of planets within $20 < P < 200$ days and $1 < R_p < 4 R_\oplus$ is $46 \pm 3\%$. This represents the average number of planets per solar-type star in the Kepler field.

4. Extrapolating our radius and period distributions out to the habitable zone for solar-type stars, we find $n_R = 6.4^{+3.4}_{-1.1}$.

5. While our results confirm those from earlier studies, there is a discrepancy in the occurrence of planets for $2.8 < R_p < 4 R_\oplus$ planets between our work (10.5 $\pm 1.0\%$) and PHM13 (6.0 $\pm 0.5\%$). Our incorporation of radius errors and updated Huber et al. (2014) stellar parameters account for about half of this discrepancy, while the difference in raw samples account for the remainder. PHM13 includes $P < 20$ day planets into their analysis, which likely contains photoevaporated planets (see Table 2), decreasing the overall occurrence in the $2.8 < R_p < 4 R_\oplus$ bin. We claim that the increase in the occurrence of $2.8 < R_p < 4 R_\oplus$ planets in our analysis is due to the exclusion of planets altered by proximity to their host stars.

6. In a detailed comparison with PHM13 we find that using CDPP values can effectively reproduce the detection completeness found by the more sophisticated analysis of PHM13.

7. Large radius errors are present in the Kepler data, and failing to account for these properly can lead to a different radius distribution. Specifically, this tends to result in a large excess of Earth-sized planets. Increasing the size of Kepler stars by 25% increases the frequency of large planets while keeping the occurrence of small planets roughly constant. Many stellar radii in the Kepler catalog are suspected to be underestimated, and Gaia will improve these stellar radius errors and resolve this issue.

The authors thank the referee, Erik Petigura, for his insightful comments that helped to improve this manuscript. This research uses data collected by the Kepler mission, as well as the NASA Exoplanet Archive. This research is supported by an NSERC CGS M award to A.S. and an NSERC discovery grant to W.Y.Q. E.G. acknowledges support by NASA grants NNX10AQ36G and NNX11AC33G.