THE ECCENTRICITY DISTRIBUTION OF SHORT-PERIOD PLANET CANDIDATES DETECTED BY KEPLER IN OCCULTATION

Megan Shabram1, Brice-Olivier Demory2, Jessi Cisewski3, Eric B. Ford3, and Leslie Rogers4

1 Department of Astronomy and Astrophysics, Eberly College of Science, The Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA
2 Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB30HE, UK
3 Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15214, USA
4 California Institute of Technology, MC249-17, 1200 East California Boulevard, Pasadena, CA 91125, USA

Received 2015 February 14; accepted 2015 September 20; published 2016 March 23

ABSTRACT

We characterize the eccentricity distribution of a sample of ∼50 short-period planet candidates using transit and occultation measurements from NASA’s Kepler Mission. First, we evaluate the sensitivity of our hierarchical Bayesian modeling and test its robustness to model misspecification using simulated data. When analyzing actual data assuming a Rayleigh distribution for eccentricity, we find that the posterior mode for the dispersion parameter is $\sigma = 0.081^{+0.014}_{-0.003}$. We find that a two-component Gaussian mixture model for $e \cos \omega$ and $e \sin \omega$ provides a better model than either a Rayleigh or Beta distribution. Based on our favored model, we find that ∼90% of planet candidates in our sample come from a population with an eccentricity distribution characterized by a small dispersion (∼0.01), and ∼10% come from a population with a larger dispersion (∼0.22). Finally, we investigate how the eccentricity distribution correlates with selected planet and host star parameters. We find evidence that suggests systems around higher metallicity stars and planet candidates with smaller radii come from a more complex eccentricity distribution.

Key words: methods: statistical – planets and satellites: dynamical evolution and stability – planets and satellites: formation

1. INTRODUCTION

The Kepler mission has identified a sample of planet candidates detected both in transit and occultation, providing detailed orbital information, including orbital eccentricity, for a subset of systems with a wide variety of stellar host properties. However, early works on the eccentricity distribution of all Kepler objects of interest (KOIs), including those in this subset, are limited due to uncertainties in host star properties. Recent studies have focused on applying Bayesian data analysis for robust error estimation (e.g., Parviainen et al. 2013), and other studies have investigated the eccentricity distribution of planets discovered with the radial velocity technique and the role that tidal interactions play in shaping eccentricity distributions (Matsumura et al. 2008; Wang & Ford 2011; Hansen & Murray 2015). While some studies have attempted to constrain the eccentricity distribution of planets via transit durations identified by Kepler, these studies have been limited by uncertainties in stellar densities (Moorhead et al. 2011; Kane et al. 2012; Plavchan et al. 2014; Van Eylen & Albrecht 2015). Lucy (2013) used a Bayesian approach to explore the eccentricity distribution of eclipsing binaries (EBs). Kipping (2014) explored biases in an eccentricity distribution using a Beta distribution prior, but little else has been done to explore the eccentricity distribution of exoplanets via similar methods and with the goal of quantifying population-level parameters. Hogg et al. (2010) proposed using a hierarchical Bayesian (HB) model to constrain the eccentricity distribution of hot Jupiters, but applied their model to simulated radial velocity observations, only.

Bayesian inference has made its way into exoplanet studies as computing facilities have evolved to accommodate the required calculations. The application of HB modeling is highly relevant for studying the Kepler planet sample (e.g., Demory 2014; Foreman-Mackey et al. 2014; Rogers 2015; Wolfgang & Lopez 2015). This framework allows us to obtain population-level posterior distributions, such as the distribution function for planets, while accounting for measurement uncertainties and potentially, selection effects. HB is particularly well suited for characterizing a population’s eccentricity distribution largely because of its ability to accommodate samples where each measurement has a large measurement uncertainty.

As a first step in studying the exoplanet population in general, we use HB modeling to investigate the eccentricity distribution of the subset Kepler planet candidates that are detected in both transit and occultation, which provides measurements of projected eccentricity via transit duration ratios and phase offsets. Even for this subset of planet candidates, individual eccentricity measurements often have large uncertainties. Fittingly, HB is designed to account for individual measurement uncertainties. Thus, we approach this problem from both sides: we will apply modern statistical methods that incorporate uncertainties into our eccentricity study (e.g., HB modeling), while also working with a subset of planet candidates with enough information to help bypass some of the uncertainty in their host star parameters. The projected eccentricity measurements ($e \cos \omega$ and $e \sin \omega$) are presumed to be independent of the stellar host star density and radius, which mitigates the problem of uncertainties in stellar parameters. Applying HB to the eccentricity distribution is a logical starting point while working to construct a comprehensive hierarchical model (i.e., a joint population distribution that includes planet parameters in addition to orbital eccentricity) in which measurement uncertainties are naturally incorporated into the analysis.

Furthermore, we can investigate various sub-populations of planets from the Kepler sample and look for correlations of planet and host star properties within these subpopulations. In
particular, we explore mixture models where the eccentricity distribution can be interpreted as a combination of two sub-populations. With this analysis, more than one population in the eccentricity distribution could arise, for example, due to different formation mechanisms at work. Characterizing the eccentricity distribution in this way provides insight into postulated planet formation theories such as planet orbital migration and planet–planet scattering. In principle, these mechanisms could form two populations that make up the eccentricity distribution: one population that evolved via slow disk migration and another population that evolved via excitation of a large eccentricity (e.g., planet–planet scattering or secular perturbations) proceeded by tidal circularization. With this in mind, the population of planets that came from planet–planet scattering might have a larger dispersion, as it would include planets with large eccentricities, while the population of planets that came from disk migration might have a smaller dispersion and contain fewer eccentric planets. These populations might also correlate with host star properties, which would allow for a framework to test physical models of the origin of each population, thus shedding light onto planet formation.

Here, we focus on inferring the eccentricity distribution of an interesting subset of planets using HB modeling applied to both simulated and real transit and occultation measurements from the Kepler mission. This sample contains predominantly short-period planet candidates, most of which are likely to be hot Jupiters, identified by Kepler. We look for correlations between the eccentricity distribution and other properties, such as stellar effective temperature, planet radius, orbital period, and stellar metallicity to begin to synthesize a global understanding of planet formation. This manuscript is organized as follows. In Section 2, we describe our observational data. In Section 3, we describe the method behind the HB analysis calculations, and the priors selected for the study. In Section 4, we present the results of our HB analysis. In Section 5, we investigate potential correlations between the eccentricity distribution and planet or host star properties. In Section 6, we summarize our results, and in Section 7, we discuss our conclusions, potential biases and future work.

2. OBSERVATIONS

When a planet both transits and occults its host star, we are able to obtain detailed information about the planetary orbit, including information about the projected orbital eccentricity, \( h = e \cos \omega \) and \( k = e \sin \omega \). The relationship between orbital eccentricity and transit observables is outlined in Winn (2010), \( h \) can be derived from

\[
\Delta t_c \approx \frac{P}{2} \left[ 1 + \frac{4}{\pi} h \right],
\]

where \( \Delta t_c \) is the time between the center of the transit and the center of the occultation and \( P \) is the orbital period. \( k \) can be derived from

\[
\frac{T_{occ}}{T_{tra}} \approx \frac{1 + k}{1 - k},
\]

where \( T_{occ} \) is the occultation duration and \( T_{tra} \) is the transit duration. More precise expressions for \( h \) and \( k \) are listed in Ragozzine & Wolf (2009), in Section 2.5 and elsewhere. The original full derivations of these expressions can be found in Sterne (1940) and de Kort (1954). When analyzing Kepler observations in Section 5, we calculate the transit and occultation times and durations numerically using Keplerian orbits.

We have measured the offsets and durations of transits and occultations for a sample of planet candidates observed by Kepler. This study is based on quarters Q0 through Q12 of Kepler data (see Burke et al. 2014, for the Q0–Q10 data release). In total, the data sets encompass about \( \sim 1100 \) days of quasi-continuous photometric monitoring between 2009 May and 2012 March. We retrieved the Q0–Q12 FITS files from MAST\(^5\) and extracted the SAPFLUX, commonly known as “calibrated light curves,” long-cadence photometry (Jenkins et al. 2010) for each target. Using the calibrated data eliminates the potential instrumental corrections or cotrending basis vectors to introduce noise correlated on short timescales.

2.1. Derivation of the Planet Physical and Orbital Properties

We focus on a sample of planet candidates for which an occultation is detected. We address the biases that the selection effects introduce into our sample in Sections 3.2 and 7. We use the Kepler planet candidate list to keep all planets larger than 8 Earth radii and with orbital periods less than 10 days. This initial selection of planet candidates was based on early planet candidate lists from NExScI. Note that most planets with a detectable occultation have very high SNR transits, so we do not expect that many additional planets would be found in the full Q0–Q17 data sets. The preliminary parameters were derived using the Kepler Input Catalog (KIC) stellar values, and updated later in our analysis. We employ a Markov Chain Monte Carlo (MCMC) framework to compute the posterior distribution of the system’s orbital parameters using these initial values. When performing MCMC analysis, we used an empirical main sequence mass–radius relationship (Torres et al. 2009) to derive more accurate planetary parameters. After the MCMC analysis was performed, some planet candidate radii changed to be outside the initial range stated above. Our MCMC implementation (described in Gillon et al. 2012) uses the Gibbs sampler and the Metropolis–Hastings algorithm to estimate the posterior distribution function of all unknown parameters. Our nominal model is based on a star and a single transiting planet on a Keplerian orbit about their centers of mass.

The input data provided to each MCMC run consist of the Q0–Q12 Kepler photometry and the stellar parameters (effective temperature \( T_{eff} \), metallicity [Fe/H] and spectroscopic log \( g \) extracted from the KIC (Brown et al. 2011). We correct for the photometric dilution induced by neighboring stellar sources using a quarter-dependent dilution factor based on the dilution values presented in the literature and on the contamination values reported in the FITS files headers (Bryson et al. 2013).

We divide the total light curve in segments of duration \( \sim 24–48 \) hr. The smooth photometric variations due to stellar variability or instrumental systematic effects in each segment are fit with a time-dependent quadratic polynomial. Baseline polynomial coefficients are determined at each step of the MCMC for each light curve with the singular value decomposition method. The resulting coefficients are then used to correct the raw photometric light curves. We assume a quadratic law for the limb-darkening (LD) and use

\[^{5}\text{http://archive.stsci.edu/kepler/}\]
c₁ = 2u₁ + u₂ and c₂ = u₁ - 2u₂ as jump parameters, where u₁ and u₂ are the quadratic coefficients (Mandel & Agol 2002). We integrate over the 29.4 minute long cadence integration time when modeling long cadence light curves.

The MCMC has the following set of jump parameters (i.e., parameters that are not fixed in our model and are used as a basis for proposal steps): the planet/star flux ratio, the impact parameter b, the transit duration from first to fourth contact, the time of minimum light T₀, the orbital period, the occultation depth, the two LD combinations c₁ and c₂ and the two parameters √e cos ω and √e sin ω. At each step of the MCMC, the Keplerian model is constructed based on the e and ω values derived from the √e cos ω and √e sin ω jump parameters. A uniform prior distribution is assumed for all jump parameters except c₁ and c₂. This corresponds to a prior that is uniform in e ∈ [0, 1) and ω ∈ [0, 2π). For the LD parameters, we assume normal priors that are centered on values of c₁ and c₂ that correspond to the values of u₁ and u₂ from the theoretical tables of Claret & Bloemen (2011) for the stellar parameters obtained from the KIC. The standard deviations of the priors for c₁ and c₂ were set by the corresponding standard deviations propagated from u₁ and u₂’s uncertainties. We run two Markov chains of 100,000 steps for each planet candidate. The mixing and convergence of the Markov chains are assessed using the Gelman–Rubin statistic criterion (Gelman & Rubin 1992). Results for e cos ω and e sin ω are shown in Table 1.

2.2. Properties of Planet Candidates Analyzed

When selecting the initial planet candidates that we perform MCMC fits for planet properties described above, we vet for EBs using the procedure outlined in Demory & Seager (2011). This leaves us with a sample of 85 planet candidates for which we have calculated posteriors for their orbital and physical properties. From this new list of planet candidates with updated properties from MCMC fitting, we do a second updated sweep for EBs referring to Tenenbaum et al. (2014) and Bryson et al. (2013), works that were published after our initial planet candidate list was developed. We also reference the Kepler EB catalog5 for additional newly reported EBs. From this procedure, we are able to exclude an additional 18 planet candidates. We include KOI 1227 in our sample of planet candidates, as it appears in both the Kepler and the two planet candidates that have approximately Gaussian measurement uncertainties for h and k to use for our analysis of the eccentricity distribution in Section 5. Working in h and k space instead of eccentricity space greatly simplifies our HB model for the eccentricity distribution (see Section 3.1), since the measurement uncertainties for h and k can be assumed to be roughly normally distributed.

The 50 remaining planet candidates have radii estimates of ~1.9–30 Earth radii, with a median value of 10.6 Earth radii, host star effective temperature of 3948–8848 K, with a median value of 5728 K, orbital period of 1.03–20.13 days with a median value of 4.24 days, and host star metallicity of −0.518 to 0.440 in [Fe/H] with a median value of 0.023 [Fe/H]. The 30 Earth Radii planet candidate (KOI 1793) is large, and is an outlier for typical radii in our sample, but still makes it past our EB vetting procedure outlined above. These values came from the Kepler Star Properties Catalog as reported from the Exoplanet Archive updated 2013 December and revised 2014 February (Buchhave et al. 2012; Huber et al. 2014). The majority of the planets have stellar metallicity values obtained from photometry, and 10 planets have spectroscopically derived stellar parameters.

3. METHOD

We aim to simulate and characterize the eccentricity distribution of a subset of the population of planet candidates in the Kepler sample for which both transits and occultations have been observed. First, we describe a general HB model, before specializing it for our application of characterizing the eccentricity distribution in Section 3.2. Next, we build and test a model using simulated data in order to determine the accuracy of our method and robustness to model misspecification in Section 4.2, then we apply our model to the real data set in Section 5.

3.1. The HB Model

HB modeling is a powerful method to estimate population parameters by propagating the unique uncertainty from each measurement of the population constituents into the inference of the population parameters. An HB model requires an analysis model that parameterizes the functional form of the population distribution, p(xₚ | φ), where xₚ represents the true value of each quantity being measured (later we adapt this model so that xₚ represents h and k). φ is the set of hyperparameters that determine the features or shape of the prescribed analysis model. To infer these population hyperparameters, we must specify the priors for the hyperparameters or the hyperpriors, p(φ). Once this multi-level model is applied to a sample of measurements, both the population’s parameters and the true parameters for each of the population members can be inferred simultaneously. The measured properties (dₚ) are related to the true properties (xₚ) and the measurement uncertainties (σₚ) by p(dₚ | xₚ, σₚ). As a result, the HB model allows us to characterize the true parameter values and population level parameters while using the information contained in the measurements and their uncertainties.

The general form for the posterior for the hyperparameter vector (φ), where D represents the number of measurements that make up the data set (dₚ), is given by

\[ p(\phi | x_p, \sigma_p) \propto p(\phi) \prod_{p=1}^{D} \left[ \int dx_p p(x_p | \phi) p(d_p | x_p, \sigma_p) \right]. \tag{3} \]

Next, let us consider a simplified HB model where each measurement dₚ is drawn from a normal distribution centered on the true value xₚ with measurement uncertainty σₚ,

\[ p(d_p | x_p, \sigma_p) \sim \text{Normal}_d(x_p, \sigma_p^2). \tag{4} \]

Here, the “~” can be read as “is distributed as,” a common notation for statisticians. At the “mid-level” of the hierarchical model, we assume that the population of true values, xₚ’s, can be parameterized by a Gaussian mixture model, where each component of the population model has mean zero and Nₘ is

5 http://exoplanetarchive.ipac.caltech.edu/docs/oeibin.html
the number of mixture components.

\[ p(x_i|\phi) = \sum_{i} f_i \text{Normal}_{\sigma}(0, \sigma_i^2) \tag{5} \]

Each component contributes a fraction \( f_i \) of the population, so

\[ \sum_{i} f_i = 1, \tag{6} \]

\( \phi \) then represents all of the \( f_i \) and \( \sigma_i \) values.\(^7\) If we assume a common Gaussian mixture model prior for each \( x_i \), as shown in Equation (5) and Gaussian measurement error as shown in Equation (4), then our hierarchical model can be

\( \sigma_i \) is a hyperparameter that partly describes the underlying population distribution along with \( f_i \), where \( \sigma_i \) is the measurement uncertainties of the observable quantity.

Note. Result for \( e \cos \omega \) and \( e \sin \omega \) from MCMC. See Section 2.1 for details on how these values are calculated.
mathematically described as Equation (3) adapted to our specific analysis:

\[
p\left(f_i, \sigma_i \mid d_p, \sigma_p \right) \propto p\left(f_i, \sigma_i \right) \\
\times \prod_{p=1}^{D} \int \sum_{i=1}^{N_m} \int dx_p \text{Normal}_{x_p}(0, \sigma_i^2) \text{Normal}_{d_p}(x_p, \sigma_p^2). \tag{7}
\]

Moving the integral inside the summation and exploiting the symmetry of the Gaussian distribution, we get

\[
p\left(f_i, \sigma_i \mid d_p, \sigma_p \right) \propto \prod_{p=1}^{D} \left[ \sum_{i=1}^{N_m} \int dx_p \text{Normal}_{x_p}(0, \sigma_i^2) \text{Normal}_{d_p}(x_p, \sigma_p^2) \right]. \tag{8}
\]

In Equation (9), we extend the limits of the integral to infinity in order to develop an analytic approximation to our hierarchical model that is accurate when \(\sigma_i^2 \lesssim 1\), \(\forall i \in [1, N_m]\) (i.e., allowing the underlying model to assign eccentricities \(\gtrsim 1\))

\[
p\left(f_i, \sigma_i \mid d_p, \sigma_p \right) \propto \prod_{p=1}^{D} \left[ \frac{\sum_{i=1}^{N_m} f_i \exp\left(-d_p^2/(\sigma_p^2 + \sigma_i^2)\right)}{\sqrt{2\pi (\sigma_p^2 + \sigma_i^2)}} \right]. \tag{9}
\]

We discuss how we modify this derivation when evaluating our model numerically, applied to the eccentricity distribution in Section 4.1. The posterior distribution for the hyperparameter vector is conditional on all observations. The posterior modes and credible intervals can be calculated from Equation (7) using MCMC or estimated analytically based on Equation (9). Recent applications of HB modeling applied to other Kepler observations include Morton & Winn (2014), Rogers (2015), Wolfgang & Lopez (2015), and Foreman-Mackey et al. (2014).

3.2. Applying the Hierarchical Model to Eccentricity Measurements

Next, we tailor the above model to the eccentricity distribution. The set of projected eccentricity measurements \(h\) and \(k\) for each planet candidate becomes the \(x_p\)'s described in Section 3.1. We assume that each true value of \(h\) and \(k\) is drawn from a distribution that is a mixture of \(N_m\) normal distributions (the analysis model), where each mixture component contributes a fraction \(f_i\) centered on zero, and has a standard deviation \(\sigma_i\). Thus, the hyperparameters \(\phi = \{f_1, \ldots, f_{N_m}, \sigma_1, \ldots, \sigma_{N_m}\}\) describe the underlying population’s distribution of \(h\) and \(k\)’s. Since fractions sum to one, \(f_{N_m} = 1 - \sum_{i=1}^{N_m-1} f_i\).

The values of \(h\) and \(k\) provide an alternate parameterization for the eccentricity (\(e\)), and the argument of periastron (\(\omega\)). We assume that the orientations of planetary systems’ pericenter directions (\(\omega\)) will be randomly distributed with respect to the direction toward Earth, i.e., \(\omega\) is uniform random \(0, 2\pi\). Thus, the prior probability distribution for each planet’s \(h\) and \(k\) has radial symmetry. While this is an excellent general model for planets, it is an approximation for our sample of planet candidates since (1) the geometric transit probability and occultation probability depend somewhat on \(\omega\) for eccentric orbits and (2) the detection probability of both the transit and occultation depends on the transit and occultation durations and thus the eccentricity and pericenter direction, and the occultation duration also depends on the orbital period. We will discuss these issues further in Section 7. Results of this analysis can be found in Sections 4 and 5.

3.3. Evaluating the Hierarchical Model

We sample from the posterior using MCMC. To calculate Markov chains we use the publicly available code Just Another Gibbs Sampler (JAGS; Plummer 2003). JAGS uses Gibbs sampling when possible, and otherwise reverts to standard random walk Metropolis–Hastings. We simultaneously sample from both the posterior distributions for the population parameters and the posterior predictive distributions for each observable. We compare the within-chain variance to the between-chain variance and evaluate the Gelman–Rubin (\(\hat{R}\)) ratio to test for non-convergence, and accept chains with an \(\hat{R} < 1.01\). We also look at the autocorrelation function for the Markov chains and accept cases that have a zero crossing at a lag of \(\lesssim 5\). The exact JAGS input model used in our study can be found online.8

4. RESULTS

4.1. Prior Specification

We consider three different analysis models and calculate posteriors for each using simulated data to test the accuracy and robustness of our method. The three analysis models used for \(x_p\)'s in our calculation are (i) a single Gaussian (\(N_m = 1\)), (ii) a two-component Gaussian mixture (\(N_m = 2\)), and (iii) a three-component Gaussian mixture (\(N_m = 3\)).

We use these same three models both to analyze the data and to generate simulated observations. In each model, the population parameters, also known as hyperparameters, \(\phi\) are a union of the set of dispersions for each mixture component \(\sigma_i\)'s and the set of fractions of planets associated with each of the mixture components \(f_i\)'s. In each case, each mixture component is a Gaussian centered at zero and represents a unique population.

When evaluating the eccentricity parameter space, we take our priors for \(h\) and \(k\) to be a mixture of Gaussian distributions, each with zero mean but truncated such that \(e = \sqrt{h^2 + k^2} \lesssim 1\). Following truncation, the prior for \(h\) and \(k\) is renormalized, so that the total probability integrates to unity. The truncation accounts for the selection effect of not detecting planets on hyperbolic orbits (\(e \gtrsim 1\)), as any such planets are not bound to their host systems and do not transit more than once. A Rayleigh distribution can also be parameterized as the square root of the sum of squared Normal distributions with zero mean, where the variance of each component is equivalent to the Rayleigh parameter. Thus, the prior population distribution for \(e\) is a truncated Rayleigh distribution for \(N_m = 1\) and can be visualized as a mixture of truncated Rayleigh distributions of \(N_m \geq 2\). Choosing a Rayleigh distribution for the eccentricity distribution is physically motivated by the fact that it naturally arises for exoplanets on circular orbits and subjected to a series of many normally distributed small random perturbations to its orbit. It is therefore a common distribution “shape” used for eccentricity distributions (e.g., Moorhead et al. 2011; Fabrycky et al.

8 http://www.astro.ufl.edu/~mshabram/jags_model/eccmodel.txt
The uncertainty in the phase offset of the transit is typically smaller than that of the occultation and transit duration ratio, thus the eclipse data constrain $h$ with more precision than $k$.  

Calculating a posterior probability distribution function (PDF) from a HB model also requires specifying a prior probability distribution for the population parameters ($\phi$). This is known as the hyperprior. Our hyperparameters are the dispersions $\sigma_i$’s (for each mixture component) and the associated mixture fractions. We assume a uniform prior for the dispersions of each mixture component between 0 and 1. The mixture component fractions follow a Dirichlet distribution with the concentration parameter set to 1 (e.g., no component is given special weight). This is the multidimensional generalization of the Beta distribution. The Dirichlet distribution forces the sum of the mixture component fractions to equal one.

4.2. Validating the Hierarchical Model

Since the true distribution parameters for the synthetically generated data sets are known, analyzing these simulated observations with our hierarchical model allows us to directly compare the output population parameters and the input population parameters. We are also able to test the sensitivity of the posterior to the chosen analysis model.

We expect to see variations in the ability of a given analysis model to recover the input model’s parameters. For instance, if the analysis model is the same as the model used to generate the simulated observations, then we expect to be able to recover the input population parameters, within the limits of measurement uncertainties and Monte Carlo error. However, if the analysis model is different than the model used to generate the data, there could be larger differences between the posterior predictive distributions for the eccentricity distribution and the actual distributions used to generate the data. If we can identify an analysis model that is relatively insensitive to the model that was used to generate the simulated observations, then we can increase our confidence in the robustness of the procedure when applying our hierarchical model to a real data set.

4.2.1. Generation and Analysis of Simulated Data

We are also interested in understanding the effect that the quantity and quality of the data has on our inference. It is important that we choose an analysis model that is relatively robust to model misspecification, so we can be confident when applying the HB model to real data. To accomplish this, we generate several simulated data sets varying the number of planets in the sample, the simulated measurement uncertainties, or both. For each pair of generative model and analysis model, we analyze four data sets of different qualities. We summarize each in Table 1. Data sets labeled “good” (“half”) consist of 50 (25) planets with measurement uncertainties of 0.04 and 0.08 for $h$ and $k$, respectively.9 These data sets are designed to be similar to our actual transit and occultation data set for both $h$ and $k$. Data sets labeled “better” (“best”) contain 50 (500) planets with measurement uncertainties of 0.001, and are designed to forecast the power of this method and model setup when the quantity of real data grows and the quality of data is improved upon (better measurement uncertainty).

We illustrate an example case in Figure 1, by showing cumulative distributions of $|h|$ and $|k|$. The solid black curve is the true distribution from which the simulated planet’s $h$ and $k$ values are drawn. The dashed black curve is the cumulative distribution for one simulated data set (“R2,” “good”; $f_1 = 0.7$,

9 The uncertainty in the phase offset of the transit is typically smaller than that of the occultation and transit duration ratio, thus the eclipse data constrain $h$ with more precision than $k$.  

### Table 2  
Model Data Sets

| Model Name | $N_p$ | $\sigma_h$ | $\sigma_k$ |
|------------|-------|------------|------------|
| Half       | 25    | 0.040      | 0.080      |
| Good       | 50    | 0.040      | 0.080      |
| Better     | 50    | 0.001      | 0.001      |
| Best       | 500   | 0.001      | 0.001      |

Note. Values indicating the quantity and quality of the suite of simulated observations used in our analysis. Data sets labeled “good” (“half”) consist of 50 (25) planets with measurement uncertainties of 0.04 and 0.08 for $h$ and $k$ respectively. These data sets are designed to be similar to our actual transit and occultation data set for both $h$ and $k$. Data sets labeled “better” (“best”) contain 50 (500) planets with measurement uncertainties of 0.001, and are designed to forecast the power of this method and model setup when the quantity of real data grows and the quality of data is improved upon (better measurement uncertainty).

First, we validate our HB model using the same model for the analysis as used to generate a simulated data set. Next, we consider the results of applying an analysis model that differs from the model used to generate the data. The purpose of making these comparisons is to identify an appropriate analysis model, balancing the need for flexibility with the desire to minimize model parameters. By analyzing a variety of simulated data sets, we develop an intuition for how different models perform, prior to analyzing the actual data. Since the true distribution of exoplanet eccentricities likely differs from any of our analysis models, it is important to analyze data sets generated under alternative models, so as to test the robustness of our approach.

When we use the same analysis model and generative model, it would be possible to compare the true model parameters to the posterior distribution for the model parameters. However, most of our comparisons involve different analysis and generative models. In these cases, it is not possible to compare the true model parameters to the posterior distribution for the model parameters. Instead, we compare the predictive posterior distribution for the eccentricity distribution (i.e., the distribution of interest). We use the Kolmogorov–Smirnov (K–S) distance to measure how the predictive posterior distribution for eccentricities under each analysis model compares to the true eccentricity distribution used to generate the simulated data. Table 3 shows the median K–S distance between each simulated data set’s true $h$ and $k$ distribution and the posterior predictive distribution for $h$ and $k$ based on 20 simulations of a particular hierarchical model from Table 2 (see Section 4.2.1 for a list of the chosen “true” eccentricity distribution values used in our study).

We illustrate an example case in Figure 1, by showing cumulative distributions of $|h|$ and $|k|$. The solid black curve is the true distribution from which the simulated planet’s $h$ and $k$ values are drawn. The dashed black curve is the cumulative distribution for one simulated data set (“R2,” “good”; $f_1 = 0.7$, 

Table 2: Model Data Sets

| Model Name | $N_p$ | $\sigma_h$ | $\sigma_k$ |
|------------|-------|------------|------------|
| Half       | 25    | 0.040      | 0.080      |
| Good       | 50    | 0.040      | 0.080      |
| Better     | 50    | 0.001      | 0.001      |
| Best       | 500   | 0.001      | 0.001      |

Note. Values indicating the quantity and quality of the suite of simulated observations used in our analysis. Data sets labeled “good” (“half”) consist of 50 (25) planets with measurement uncertainties of 0.04 and 0.08 for $h$ and $k$ respectively. These data sets are designed to be similar to our actual transit and occultation data set for both $h$ and $k$. Data sets labeled “better” (“best”) contain 50 (500) planets with measurement uncertainties of 0.001, and are designed to forecast the power of this method and model setup when the quantity of real data grows and the quality of data is improved upon (better measurement uncertainty).
If the model is working properly, we expect to get posterior distributions for the population’s parameters that are consistent with the values used to generate the data. Indeed, we find K–S distances are between ~0.05 and 0.1 for these cases. Since models have different parameters (even when they are represented by the same variable names), the most appropriate way to compare the performance is based on the posterior predictive distribution for the population of measurements. In this case, the K–S distances between the posterior predictive distributions for the HB model and actual model are ~0.1 to 0.2.

For simulated data sets with smaller measurement uncertainties, we find that the K–S distance between the posterior predictive distribution and associated simulated data for $h$ and $k$ is similar for analysis models that have at least the same number for mixture components or more. Additionally, for several combinations of analysis and generated models, we note that the $N_m = 2$ analysis model results in a smaller K–S distance to the R3 data than the $N_m = 3$ analysis model. This is likely due to the greater flexibility of the $N_m = 3$ model and finite number of measurements, i.e., the three-component model “over-fits” the discrete data set. We found that the $N_m = 2$ analysis model did a better job overall at recovering the predictive distribution for the simulated data sets across all versions of simulated data.

Some authors advocate parameterizing the eccentricity distribution as a Beta distribution, $e \sim \text{Beta}(\alpha, \beta)$ (e.g., Kipping 2014). Therefore, we also investigate using a Beta distribution analysis model using one (“R2,” “good”) simulated data set (see Table 2). In this model set up, $\alpha$ and $\beta$ become the hyperparameters (population level parameters) that we wish to infer. We use a Gamma distribution with $\kappa = 2$ and $\theta = 1$ as the prior probability distribution for $\alpha$ and $\beta$. Figure 2 shows the results of this HB model as eccentricity versus cumulative fraction. The simulated eccentricity data are shown as the dotted black curve. The true eccentricity distribution generated using a two-component Gaussian mixture model for $h$ and $k$ (e.g., $f_1 = 0.7, f_2 = 0.3, \sigma_1 = 0.05$, and $\sigma_2 = 0.3$, as described in Section 4.2.1) is shown in red. The dashed green curve is plotted using the posterior modes for $\alpha$ and $\beta$ ($\alpha = 0.11_{-0.02}^{+0.04}$, $\beta = 1.73_{-0.82}^{+0.84}$) for this HB model. The K–S distance between the “R2,” “good” distribution (red) used and the distribution using the posteriors modes of $\alpha$ and $\beta$ (dashed-green) is 0.5, which is in support of these being two distinct distributions. Our results indicate that the standard Beta distribution is a poor choice for an analysis model to parameterize the eccentricity distribution as it erroneously predicts a strong peak near $e = 0$ and underpredicts the frequency of larger eccentricities.

Table 3

| Generalized Model Name | Analysis Model |
|------------------------|----------------|
|                        | $(N_m = 1)$     | $(N_m = 2)$     | $(N_m = 3)$     |
| R1 half                | 0.1310          | 0.1615          | 0.1740          |
| R1 good                | 0.1550          | 0.1650          | 0.1700          |
| R1 better              | 0.0990          | 0.1150          | 0.1275          |
| R1 best                | 0.0400          | 0.0500          | 0.0550          |
| R2 half                | 0.2365          | 0.2100          | 0.2715          |
| R2 good                | 0.2545          | 0.1875          | 0.2330          |
| R2 better              | 0.1890          | 0.0825          | 0.1590          |
| R2 best                | 0.1705          | 0.0660          | 0.1485          |
| R3 half                | 0.1535          | 0.1215          | 0.1415          |
| R3 good                | 0.2560          | 0.2085          | 0.2285          |
| R3 better              | 0.1270          | 0.1035          | 0.1350          |
| R3 best                | 0.1726          | 0.0835          | 0.2142          |

Note. Results of validation and sensitivity analysis of three hierarchical models for eccentricities. Shown in this table are Kolmogorov–Smirnov (K–S) statistics comparing data sets of simulated observations with data sets generated using the posteriors of the hyperparameters from an HB model that analyzed the same set of simulated observations (comparing input to output to test model). Here R1, R2, and R3 represent a one-, two- and three-component Gaussian mixture model, respectively. Table 2 summarizes the different quantity and quality of simulated observations used in this analysis. A two-component Gaussian mixture model does well across the majority of simulated data sets. See Section 4.2.2 for a detailed interpretation of these results.
and model for curve. The distribution generated using a two-component Gaussian mixture distribution of the simulated eccentricity data is shown as the dotted black curve. The distribution erroneously predicts a strong peak near frequency of larger eccentricities.

We investigate using a Beta distribution analysis Figure 2. Results of an HB model that parameterizes the eccentricity as a standard Beta distribution. We calculated posteriors of one-, two-, and three-component Gaussian mixture models applied to real Kepler transit and occultation data set. The 68.3% credible intervals about the mode are shown as dotted black lines, and the mode is shown as a vertical solid black line. A one-component Gaussian model is insufficient at characterizing the eccentricity distribution of our sample of planet candidates from Kepler. Shabram et al., 2016 April 1

In principle, we could model the transit and occultation times and durations directly, instead of $h$ and $k$. However, this would increase the computational complexity. Even taking advantage of our approximations, the calculations presented represent a significant computational investment (~2.4 CPU months). Modeling $h$ and $k$ also facilitates deriving analytical expressions for testing the algorithms. By modeling the projected eccentricity, $h$ and $k$, we were able to thoroughly test both the code and algorithm on real and simulated data sets.

5. RESULTS FOR KEPLER PLANET CANDIDATES WITH OCCULTATIONS

We calculated posteriors of one-, two-, and three-component Gaussian mixture models applied to real Kepler transit and occultation data set. Figure 3 shows a histogram of the observed $h$ and $k$ values from Table 1 (shown in gray). Since we are assuming the argument of periastron ($\omega$) is random, $h$ and $k$ are equivalent, or drawn from the same distribution. A Gaussian distribution using population parameters from the posterior mode for the dispersion for a one-component model is overplotted (shown as the dotted black curve). This one-component model does a poor job at capturing the shape of the distribution because it struggles to match the moderate eccentricity outliers. A two-component Gaussian mixture model using population parameters from the mode of the 2D marginal posterior for the mixture fraction and dispersions is shown in red. This model captures the peaked nature of the true distribution while also allowing for a smaller number of measurements far from the central peak.

Figure 3. A histogram of the $h$ and $k$ data set is shown in gray. The dotted black curve is a one-component Gaussian distribution using the posterior mode for the dispersion obtained from an HB model that uses a one-component Gaussian mixture model. Shown in red is a two-component Gaussian mixture model using posterior modes for the mixture fractions and dispersions obtained from an HB model that uses a two-component Gaussian mixture model. The black model does a poor job at capturing the shape of the distribution. The red model captures the peaked nature of the true distribution while also allowing for a smaller number of measurements far from the central peak.

Figure 4. Posterior distribution for the dispersion of a Gaussian model for $h$ and $k$ applied to our full Kepler short-period planet candidate transit and occultation data set. The 68.3% credible intervals about the mode are shown as dotted black lines, and the mode is shown as a vertical solid black line. A one-component Gaussian mixture model is insufficient at characterizing the eccentricity distribution of our sample of planet candidates from Kepler.

find that only two of the three components can be constrained given the quantity and quality of the Kepler transit and occultation data set, suggesting that the available data are not able to indicate the presence of a third population, or that a third population may not exist.

The posterior distribution for the dispersion of true values of $h$ and $k$ assuming a one-component Gaussian model is displayed in Figure 4, which is based on our full data set for
planet candidates with both Kepler transit and occultation measurements. The posterior mode for the dispersion is $\sigma = 0.081^{+0.003}_{-0.014}$. We use this value as the dispersion for our one-component Gaussian population model shown as the dotted black curve in Figure 3.

Next, we investigate joint posterior distributions for a two-component Gaussian mixture model applied to our full Kepler transit and occultation data set. These results are shown in Figure 5, where the panels on the diagonal show the marginalized posterior distribution for the population parameters: $\sigma_{\text{low}}$ the lesser value of $\sigma_1$ and $\sigma_2$, $\sigma_{\text{high}}$ the greater value of $\sigma_1$ and $\sigma_2$, and $f_{\text{low}}$, the weights for the mixture component ($f_{\text{high}} = 1 - f_{\text{low}}$). The use of $\sigma_{\text{low}}$ and $\sigma_{\text{high}}$ instead of $\sigma_1$ and $\sigma_2$ (and corresponding fractions) is helpful for visualizing the results, since our model has symmetry under exchanging $(\sigma_1, f_1)$ and $(\sigma_2, f_2)$. The off-diagonal panels show posterior samples and contours for the 68.3% credible interval of the two-dimensional marginal posteriors for each parameter pair. The fact that $(\sigma_{\text{low}}, f_{\text{low}})$ and $(\sigma_{\text{high}}, f_{\text{high}})$ form two distinct clusters demonstrates the value of a two-component (two population) model for the eccentricity distribution of our sample of Kepler planet candidates.

As expected, the uncertainties in measurements of $k = e \sin \omega$ are much greater than the uncertainties in the measurements of $h = e \cos \omega$. We note that in our sample, the $k$ values are more tightly clustered around zero than $h$. Therefore, we investigated if excluding these values significantly impacts our results. By doing this we are decreasing our effective sample size, but maintaining the number of measurements with small uncertainties. When running our simulations without $k$ values, we get the following results: using a single Gaussian model $\sigma = 0.074^{+0.016}_{-0.003}$, and when using a two-component Gaussian mixture model, the marginal posterior modes for the mixture fractions and dispersions are $f_{\text{low}} = 0.93^{+0.029}_{-0.051}$, $\sigma_{\text{low}} = 0.003^{+0.001}_{-0.003}$, and $f_{\text{high}} = 0.07^{+0.062}_{-0.019}$, $\sigma_{\text{high}} = 0.187^{+0.547}_{-0.028}$, respectively. These values differ from the values obtained using the full $h$ and $k$ planet candidate data set by $9.0\%$, $4.4\%$, $107.7\%$, $44.4\%$, and $16.2\%$ for $\sigma$, $f_{\text{low}}$, $\sigma_{\text{low}}$, $f_{\text{high}}$, and $\sigma_{\text{high}}$, respectively. Each of these overlaps with the 68.3% credible interval for the same parameters with the full data set. The most notable difference is for $\sigma_{\text{low}}$, which we estimate to be 0.03 when using $k$ alone, but 0.01 when including both $h$ and $k$ observations.

5.1. Correlation of The Eccentricity Distribution with Star and Planet Properties

Another goal of this study is to investigate potential correlations between the eccentricity distribution and planet properties.

Figure 5. Joint posterior distributions for a two-component Gaussian mixture model applied to our Kepler short-period planet candidate transit and occultation data set. In each panel, the data are plotted with the horizontal axis representing $\sigma_{\text{low}}$ the lesser value of $\sigma_1$ and $\sigma_2$, $\sigma_{\text{high}}$ the greater value of $\sigma_1$ and $\sigma_2$, and $f_{\text{low}}$ the corresponding weight for the low mixture component, on a logarithmic scale. Since $f_{\text{high}} = 1 - f_{\text{low}}$, we only show $f_{\text{low}}$ here. The vertical axis shows the same variables, and each panel is the corresponding two-dimensional marginal posteriors for each parameter pair. The contour region plotted over the sampled posterior represents the 68.3% credible interval. The one-dimensional histograms are plotted as log density, with the 68.3% credible intervals shown as dotted black lines, and the mode is shown as a vertical solid black line. The two-component Gaussian mixture model characterizes the eccentricity distribution of our sample of planet candidates better than a one-component analysis model.
or host star properties. Specifically, we consider whether the planet candidate eccentricity distribution is correlated with stellar metallicity, host star effective temperature, planet radius, or orbital period. Values for each planet candidate are obtained from the Kepler Star Properties Catalog as reported in the Exoplanet Archive updated 2013 December and revised 2014 February (Buchhave et al. 2012; Huber et al. 2014). The majority of the planets have effective temperatures and stellar metallicity values obtained via KIC photometry, however, 10 planets have spectroscopically derived values. Given the relatively small sample size, we focus on comparing the distribution of planet candidates with large and small values for each parameter. We sort the planet candidates in our data set from largest to smallest values of a given property, and then create two sub-samples of the original population. Unless otherwise specified, we divide the data in half to maximize the statistical power when comparing the two samples and to avoid introducing an additional parameter specifying the dividing point between the high and low subsets. We analyze each subset as described in Section 3.

Initially, we evaluate each subset of data using an HB model with a one-component Gaussian distribution for $h$ and $k$, as we did before for the full data set. This is shown in Figure 6, where we have applied the HB model to the small-valued (blue) and large-valued (red) halves of the Kepler occultation data, sorted by (a) stellar effective temperature, (b) planet radius, (c) orbital period, and (d) stellar metallicity. The dotted lines correspond to the 68.3% credible intervals and the solid vertical lines correspond to the mode for each posterior distribution. For panel (b) planet radius, and panel (d) stellar metallicity, differences in the small-values and large-valued data subsets merit further investigation. In order to explore these results further, we analyze these subset using a two-component Gaussian mixture model (see Figure 6).

Next, we look at posterior distributions based on applying the HB model using a two-component Gaussian mixture for the analysis model applied to the small-valued half or large-valued half of the data subsets, again based on sorting by (a) stellar effective temperature, (b) planet radius, (c) orbital period, and (d) stellar metallicity. Figure 7 shows posterior distributions for the small-valued half of data (blue and green clusters), and large-valued half (red and orange clusters). The two groups of clusters represent samples of the posterior distribution for the
hyperparameter vector, in this case for $\sigma_{\text{low}}$ and $f_{\text{low}}$ (top left group of clusters in each sub-plot), and $\sigma_{\text{high}}$ and $f_{\text{high}}$ (bottom right group of clusters in each sub-plot). The data are plotted with the vertical axis representing the low value of the mixture fraction, $f_{\text{low}}$, in green and orange, and, $f_{\text{high}}$, in blue and red for the two subsets of sorted data shown. The contours represent 68.3% credible intervals. Interestingly, for planet radius (b), and stellar metallicity (d), the posteriors of the mixture fractions for the planet candidates with large-valued planet radii and for small-valued host star metallicities are consistent with 0 and 1, indicating only one population is required to accurately model the eccentricity distribution for these subsets of planet candidates. For planet radius and stellar metallicity, we also see that the two-component population models are somewhat different for small and large value subsets.

![Figure 7. Two-component Gaussian mixture model applied to the Kepler transit and occultation data. We apply an HB model to small- and large-valued halves of the short period Kepler candidate occultation data, sorted by stellar effective temperature (a), planet radius (b), orbital period (c), and stellar metallicity (d). The full sample is divided into two equally sized small- and large-value subsets before being processed through our HB model. The small-valued subset of data corresponds to the blue and green clusters, and the large-valued subset corresponds to the red and orange clusters. The two groups of clusters represent samples of the posterior distribution for the hyperparameter vector, in this case for $\sigma_{\text{low}}$ and $f_{\text{low}}$ (top left group of clusters), and $\sigma_{\text{high}}$ and $f_{\text{high}}$ (bottom right group of clusters). The data are plotted with the vertical axis representing the low value of the mixture fraction, $f_{\text{low}}$, in green and orange, and, $f_{\text{high}}$, in blue and red for the two subsets of sorted data shown. The contours represent 68.3% credible intervals. Interestingly, for planet radius (b), and stellar metallicity (d), the posteriors of the mixture fractions for the planet candidates with large-valued planet radii and for small-valued host star metallicities are consistent with 0 and 1, indicating only one population is required to accurately model the eccentricity distribution for these subsets of planet candidates. For planet radius and stellar metallicity, we also see that the two-component population models are somewhat different for small and large value subsets.](image-url)

The Astrophysical Journal, 820:93 (15pp), 2016 April 1

Shabram et al.
We modeled the eccentricity distribution using transit and occultation data. Specifically, we used the HB (Hierarchical Bayes) formulation, which is well-suited for characterizing the eccentricity distribution of planets detected in both transit and occultation. We demonstrated that the present data does not allow us to empirically identify a period cutoff that depicts two populations.

By dividing the data into equal sized subsets, we place a uniform prior on the period break point. Figure 8 shows the marginal posterior distribution for the period break point from analyzing the actual data set. The marginal posterior for the period break has peak values that are clustered near the minimum and maximum of the period values in our sample. This indicates that the model favors period breaks causing one data subset to have so few observations that the population parameters from one subset are minimally constrained. We conclude that the present data set is not able to usefully constrain this more complex model. Therefore, we do not attempt to apply a similar model allowing for two eccentricity distributions with an unknown break in terms of the host star metallicity or planet radius, since their measurement uncertainties are much larger.

6. SUMMARY OF RESULTS

We investigated the eccentricity distribution for a sample of short-period single-planet candidate systems from Kepler that are detected in both transit and occultation. We demonstrated that HB models are well-suited for characterizing the eccentricity distribution using transit and occultation data. We modeled the distribution of $h = e \cos \omega$ and $k = e \sin \omega$ measurements, derived from transit and occultation measurements. If we model the population distribution of $h$ and $k$ with a single Gaussian, then we infer a dispersion of $\sigma = 0.081 \pm 0.014$. When we applied the two-component mixture model to the full data set, we found $f_{\text{low}} = 0.89 \pm 0.057$, $\sigma_{\text{low}} = 0.01 \pm 0.014$ and $f_{\text{high}} = 0.11 \pm 0.057$, $\sigma_{\text{high}} = 0.22 \pm 0.100$. These results suggest the presence of a small population of planet candidates ($\sim 11\%$) that contain planets with a broad range of orbital eccentricities and a larger population of planet candidates ($\sim 89\%$) that contain planets on very nearly on circular orbits.

We present evidence that host stars in our sample with higher metallicity and planet candidates with smaller radii have a more complex eccentricity distribution than stars with low metallicity and planet candidates with larger radii. The eccentricity distributions of these more complex populations are well described by a two-component Gaussian mixture model with a zero mean, suggesting a potential physical explanation in terms of proposed planet formation models, which we will describe in more detail in Section 7.

7. DISCUSSION

Previous studies of the period-size distribution of Kepler planet candidates have identified two common architectures of planetary systems: (1) Systems with Tightly packed Inner Planets (STIPS; Lissauer et al. 2011; Payne et al. 2013; Boley et al. 2014) and (2) systems with a single short-period planet (often a giant planet) with either no additional planets detected or a large gap between the short-period planet and the next detectable planet (Steffen et al. 2012a, 2013; Dawson & Murray-Clay 2013). This work focuses on a sample containing primarily isolated giant planet candidates. Of course, these systems may have undetected companions, particularly at larger orbital separations where the geometric transit probability is small.

Two broad classes of mechanisms have been proposed to explain the formation of hot Jupiters. In both models, planets form at greater distances from their host star. In one model, a giant planet experiences a gradual inspiral through a gas or planetesimal disk until they halt near their present orbit (e.g., Kley & Nelson 2012). In the other model, gravitational perturbations from another massive body (potentially another planet or a stellar companion) excite a giant planet’s orbital eccentricity until its periastron distance is so small that tidal forces begin to circularize the orbit (e.g., Rasio & Ford 1996; Fabrycky & Tremaine 2007; Naoz et al. 2011). Several studies have assessed the relative merit of these two classes of models, making use of the observed orbital period distribution (Ford & Rasio 2006; Valsecchi & Rasio 2014), spin–orbit obliquity

Figure 8. Marginal posterior distribution for the critical period break point from a joint period-eccentricity distribution HB model. We analyze the full data set using an HB model that allows for the eccentricity distribution to differ depending on whether the orbital period is greater or less than the critical period break point. We infer that the present data does not allow us to empirically identify a period cutoff that depicts two populations.

model in our HB model. Our results indicate that the standard Beta distribution is a poor choice for an analysis model to parameterize the eccentricity distribution.

Next, we applied HB modeling to analyze a real data set of $h = e \cos \omega$ and $k = e \sin \omega$ measurements, derived from transit and occultation measurements. If we model the population distribution of $h$ and $k$ with a single Gaussian, then we infer a dispersion of $\sigma = 0.081 \pm 0.014$. When we applied the two-component mixture model to the full data set, we found $f_{\text{low}} = 0.89 \pm 0.057$, $\sigma_{\text{low}} = 0.01 \pm 0.014$ and $f_{\text{high}} = 0.11 \pm 0.057$, $\sigma_{\text{high}} = 0.22 \pm 0.100$. These results suggest the presence of a small population of planet candidates ($\sim 11\%$) that contain planets with a broad range of orbital eccentricities and a larger population of planet candidates ($\sim 89\%$) that contain planets on very nearly on circular orbits.

Next, we assessed whether there is evidence for more complexity in the eccentricity distribution by considering analysis models that allow correlations between the eccentricity distribution and other planet or host star parameters. For the current sample of Kepler planet candidates seen in both transit and occultation, we find interesting correlations of the eccentricity distribution with either the planet radius or the host star metallicity, but not with stellar effective temperature or orbital period.

We present evidence that host stars in our sample with higher metallicity and planet candidates with smaller radii have a more complex eccentricity distribution than stars with low metallicity and planet candidates with larger radii. The eccentricity distributions of these more complex populations are well described by a two-component Gaussian mixture model with a zero mean, suggesting a potential physical explanation in terms of proposed planet formation models, which we will describe in more detail in Section 7.
distribution (Morton & Johnson 2011; Albrecht et al. 2012; Naoz et al. 2012; Dawson 2014) and orbital architectures (Dawson et al. 2012; Steffen et al. 2012b; Dawson & Murray-Clay 2013). Collectively, these studies suggest that multiple mechanisms likely contribute to the formation of hot Jupiters. In this case, the two populations would likely have different distributions of orbital eccentricities, with disk migration leading to the smaller dispersion of eccentricities. This motivates us to consider interpreting evidence of a two-component mixture model for the eccentricity distribution in terms of two formation models.

Our analysis of Kepler’s short-period planet candidates with occultation measurements suggests that the eccentricity distribution can be well described by a two-component mixture model, where the less abundant population of planets has a broader dispersion of eccentricities. If the mixture components indeed translate into formation mechanisms of hot Jupiters, then this could suggest that disk migration could be the more common formation mechanism. Alternatively, if tidal circularization of highly eccentric proto-hot Jupiters is sufficiently rapid, then the current eccentricities could reflect late-stage excitation of orbital eccentricities due to undetected planets. Of course, a complete formation theory would need to explain all observations, including the low abundance of additional planets near hot Jupiters (Steffen et al. 2012a), the final semimajor axis of hot Jupiters (e.g., Ford & Rasio 2008; Valsecchi & Rasio 2014), the distribution of orbital obliquities (e.g., Fabrycky & Winn 2009; Morton & Johnson 2011; Albrecht et al. 2012) and correlations between obliquity and other star and planet properties (e.g., Winn 2010; Morton & Winn 2014).

It is particularly interesting to compare the eccentricity distribution of our sample to that of other subsamples of the Kepler planet candidate list, particularly subsamples dominated by smaller planets. Since the eccentricity affects the transit duration (Barnes 2007; Burke 2008), the distribution of transit durations can constrain the eccentricity distribution for arbitrary sub-samples of Kepler planet candidates (Ford et al. 2008). Early studies of the eccentricity distribution of Kepler’s planet candidates (Moorhead et al. 2011; Kane et al. 2012; Plavchan et al. 2014) were limited due to the uncertainty in stellar parameters. Transit durations combined with stellar properties from photometry (Moorhead et al. 2011), high-resolution spectroscopy (Buchhave et al. 2012; Dawson & Johnson 2012) and/or flicker (Kipping et al. 2014) can effectively recognize high eccentricity planets. However, further research is needed to obtain stellar properties that are precise and accurate enough to enable population studies of the more typical low eccentricity planets. Fortunately, one can characterize the eccentricity distribution of substantial subsets of Kepler planet candidates, either by using stars with high quality stellar characterization (e.g., asteroseismology, Huber et al. 2013; E. Ford et al. 2015, in preparation) or by using ratios that eliminate the dependence on stellar properties (e.g., Kipping 2011; Fabrycky et al. 2014; A. Morehead et al. 2015, in preparation).

Previous studies suggest that the typical eccentricity of planet candidates in systems with multiple transiting planet candidate systems (0.00–0.06) is likely smaller than in our sample (e.g., Fang & Margot 2012; Fabrycky et al. 2014). Similarly, Wu & Lithwick (2013) and Hadden & Lithwick (2014) analyze transit timing variations (TTVs) in systems with near-resonant planet candidates. Hadden & Lithwick (2014) report a maximum likelihood estimate of the dispersion of eccentricities of $\sigma_e = 0.18 \pm 0.005$. This is significantly smaller than the $\sigma = 0.08 \pm 0.003$ of our one-component model for the eccentricity distribution. Their result is comparable to the $\sigma_{\text{low}} = 0.01 \pm 0.012$ that describes nearly 90% of planets when using our two-component model. Hadden & Lithwick (2014) explicitly consider dividing their sample into subsets based on the estimated planet size being larger or smaller than $2.5R_\ast$. Interestingly, they find an even smaller eccentricity dispersion ($\sigma = 0.008 \pm 0.005$) for subset of larger planets with measurable TTVs. This is the opposite of what would be expected based on a simple comparison to either our results or planets discovered by radial velocity surveys, both of which are dominated by significantly larger and/or more massive planets. Future observations and analyses with improved statistical methodology will be important for understanding the underlying nature of these differences.

### 7.1. Potential Biases

We note several potential sources of bias in our characterization of the eccentricity distribution. First, there is a purely geometrical effect due to the fact that we analyze only planet candidates observed to both transit and occult their host stars. Since the geometrical transit and occultation probabilities are both functions of the orbital eccentricity and direction of pericenter, the eccentricity distribution of planets in our sample is different than the eccentricity distribution of all planets (even if we controlled for orbital separation relative to the star radius). Mathematically, we assume a uniform distribution for the argument of periastron, $\omega$, which is true for all planets in nature, but not true for our sample. Previous studies have suggested that the difference between the eccentricity distribution of transiting planets and all short-period planets is modest (Burke 2008; Kipping 2014). For our sample, the effect will be even weaker, since we require both transit and occultation to be included in our sample and the two have opposite dependence on $e \sin \omega$. We confirm this by computing the distribution of $\omega$ for a simulated population generated by starting with a uniform distribution and rejecting planets that do not both transit and occult. We find $\omega$ to be very nearly uniformly distributed for these cases, where there is a non-zero effect but it is less that 4%. Therefore, we do not account for this effect in our study.

We quantify the significance of this effect in an upcoming study by employing Approximate Bayesian Computing (ABC), which can naturally model complex selection criteria such as this (J. Cisewski et al. 2015, in preparation).

A second potential bias in our characterization of the eccentricity distribution is simply that the population of planet candidates we study may not be representative of all planet candidates. The decreasing geometric transit probability as a function of orbital period or semimajor axis inevitably leads to our sample being dominated by planet candidates with short orbital periods, similar to all population studies based on known transiting planets. Detection probability is also a function of the signal-to-noise of the transit and occultation. Since the transit is typically much deeper than the occultation, the detection probability for the occultation (rather than the transit) is the dominant effect for this study. The occultation signal-to-noise depends on the occultation depth and duration. In principle, the duration depends on the density of the host star, impact parameter, eccentricity, and pericenter. In practice, the occultation duration is most sensitive to the impact.
parameter. Thus, our sample may have excluded some planets with large impact parameters, resulting in the occultation going undetected. Similarly, the occultation depth depends on the effective temperature of the planet, and thus indirectly on the effective temperature of the star, the orbital distance and the stellar radius. Therefore, our sample is likely enriched in planets with larger radii, planets with higher effective temperatures, host stars with higher effective temperatures and smaller radii, and planets orbiting even more closely to their host star. If planet formation proceeds differently around more massive stars, then the eccentricity distribution for planets in our sample could deviate from the eccentricity distribution of the overall planet population.

Furthermore, our sample could be enhanced with planets with significant eccentricity if tidal forces on planets in eccentric orbits led to substantial heating and increased thermal emission. Since the sample analyzed in this study consists of mostly giant planets and host stars with a single detected transiting planet, the eccentricity distribution for our sample may differ from the eccentricity distribution of smaller planets and/or planets in systems with multiple closely spaced planets. These potential biases can also be viewed as opportunities to constrain planet formation and tidal theories. By comparing the eccentricity distribution of different planet populations, future studies can quantify how the eccentricity distribution changes with planet size, multiplicity, and stellar properties.

We anticipate several ways that future observations will allow for improvements to our analysis. First, we analyzed a subsample of the Kepler planet candidates that had already been evaluated for any indication that the KOI was actually due to an EB star, rather than a planet (e.g., Bryson et al. 2013; Tenenbaum et al. 2014). Both the transit shape and comparison of the target centroid location during and out-of-transit provide powerful diagnostics for recognizing likely false positives. Estimates of the false positive rate are sufficiently low (∼10%–20%; Fressin et al. 2013; Burke et al. 2014), that we can interpret our results in terms of the eccentricity distribution of planets. Nevertheless, one should be cognizant that the sample of planet candidates we analyze may include one or more false positives, such as diluted EBs. In particular, our study necessarily selects planet candidates for which an occultation is measured, which may lead to an increased rate of diluted EB false positives. Properly accounting for a non-zero fraction of false positives would require a significantly more complex model. Therefore, we leave such work for future studies. Alternatively, future observations of these very interesting planet candidates may identify any remaining false positives and characterize the false positive rate sufficiently well that adding further complexities to the model is not necessary.

7.2. Future Research

Transit and occultation observations from future missions could lead to improved understanding of the eccentricity distribution of short-period planets as a function of host star mass and temperature. In particular, there could be differences between a volume-limited sample of target stars and our sample due to the target selection algorithm for the Kepler planet search targets as well as variations in Kepler’s detection sensitivity. Given the Kepler target selection criteria and detection sensitivity, most of the planet candidates we analyze are orbiting F and G stars. Future missions such as NASA’s Transiting Exoplanet Survey Satellite (TESS) (Ricker et al. 2014) and ESA’s PLATO (Rauer et al. 2014) are expected to survey a broader range of target stars and to have a simpler target selection function.

Finally, future observations and analysis of host star properties could also result in improved characterization of correlations between the eccentricity distribution and host star properties. In particular, a large fraction of host star metallicities used in our analysis were derived from photometric observations as opposed to higher quality spectroscopic observations (Huber et al. 2013). As metallicities derived from high-resolution spectroscopy are published for more stars in our sample, we will be able to make more robust conclusions about the potential correlation of the eccentricity distribution with metallicity.

This paper demonstrates that it is practical to apply rigorous HB models to evaluate key dynamical properties of exoplanet populations. In principle, these methods can be readily generalized to provide a more accurate characterization of other aspects of the exoplanet population, such as the frequency of planets as a function of size and orbital period (Foreman-Mackey et al. 2014), the planet radius–mass relationship (Rogers 2015; Wolfgang & Lopez 2015), the distribution of mutual orbital inclinations and multiplicity, and the frequency of small planets in the habitable zone of solar-type stars (Foreman-Mackey et al. 2014). A challenge for future HB analysis will be to develop rigorous model comparison techniques. Each unique problem is often limited by the statistical power of the data, where there is no universal technique applicable in all cases. In practice, the high-dimensional integration required can be computationally challenging. Therefore, careful thought and problem specification is needed, so as to render the necessary calculations tractable. Fortunately, recent collaborations between astronomers and statisticians, such as the 2013 program on Modern Statistical and Computational Methods for Analysis of Kepler Data (SAMSI) at the Statistical and Applied Mathematical Sciences Institute have significantly enhanced the level of sophistication among exoplanet researchers. Forthcoming publications will describe recent efforts application of importance sampling and ABC to enable application of HB models to more complex problems (e.g., Morton & Winn 2014; J. Cisewski et al. 2015, in preparation; Rogers 2015).

We thank the entire Kepler team for the many years of work that have proven so successful and was critical to this study. We acknowledge the SAMSI Bayesian Characterization of Extrasolar Populations working group for discussions that improved this manuscript. In particular, we thank Merlise Clyde, Darin Ragozzine, Thomas Loredo, David Hogg, Thomas Barclay, and Robert Wolpert for their valuable contributions to the project. This work was supported in part by the Pennsylvania State University’s Center for Exoplanets and Habitable Worlds, NASA Kepler Participating Scientist Program awards NNX12AF73G and NNX14AN76G, and NASA Origins of Solar Systems awards NNX13AF61G and NNX14AI76G. This material was based upon work partially supported by the National Science Foundation under grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors
acknowledge the Institute for CyberScience at The Pennsylvania State University for providing advanced computing resources and services that have contributed to the research results reported in this paper (http://rcc.its.psu.edu), in particular Hoofar Pourzand and William Brouwer. The authors acknowledge the University of Florida Research Computing (http://researchcomputing.ufl.edu) for providing computational resources and support that have contributed to the research results reported in this publication. Some/all of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX13AC07G and by other grants and contracts. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. We also acknowledge Dan Foreman Mackey’s triangle.

Facility: Kepler

REFERENCES

Albrecht, S., Winn, J. N., Johnson, J. A., et al. 2012, ApJ, 757, 18
Barnes, J. W. 2007, PASP, 119, 986
Boley, A. C., Morris, M. A., & Ford, E. B. 2014, ApJL, 792, L27
Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, AJ, 142, 112
Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al. 2013, PASP, 125, 889
Buchhave, L. A., Latham, D. W., Johansen, A., et al. 2012, Natur, 486, 375
Burke, C. J. 2008, ApJ, 679, 1566
Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, ApJS, 210, 19
Claret, A., & Bloemen, S. 2011, A&A, 529, 75
Dawson, R. I. 2014, ApJL, 790, L31
Dawson, R. I., & Johnson, J. A. 2012, ApJ, 756, 122
Dawson, R. I., Johnson, J. A., Morton, T. D., et al. 2012, ApJ, 761, 163
Dawson, R. I., & Murray-Clay, R. A. 2013, ApJL, 767, L24
de Kort, J. M. A. 1954, RA, 3, 109
Demory, B.-O. 2014, ApJL, 789, L20
Demory, B.-O., & Seager, S. 2011, ApJS, 197, 12
Fabrycky, D., & Tremaine, S. 2007, ApJ, 669, 1298
Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2014, ApJ, 790, 146
Fabrycky, D. C., & Winn, J. N. 2009, ApJ, 696, 1230
Fang, J., & Margot, J.-L. 2012, ApJ, 761, 92
Ford, E. B., Quinn, S. N., & Veras, D. 2008, ApJ, 678, 1407
Ford, E. B., & Rasio, F. A. 2008, ApJ, 686, 621
Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64
Fressin, F., Torres, G., Charbonneau, D., et al. 2013, ApJ, 766, 81
Gelman, & Rubin 1992, StaSc, 7, 457
Gillon, M., Demory, B.-O., Benneke, B., et al. 2012, A&A, 539, A28
Hadden, S., & Lithwick, Y. 2014, ApJ, 787, 80
Hansen, B. M. S., & Murray, N. 2015, MNRAS, 448, 1044
Hogg, D. W., Myers, A. D., & Bovy, J. 2010, ApJ, 725, 2166
Huber, D., Chaplin, W. J., Christensen-Dalsgaard, J., et al. 2013, ApJ, 767, 127
Huber, D., Silva Aguirre, V., Matthews, J. M., et al. 2014, ApJS, 211, 2
Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al. 2010, ApJL, 713, L120
Kane, S. R., Ciardi, D. R., Gelino, D. M., & von Braun, K. 2012, MNRAS, 425, 757
Kipping, D. M. 2011, MNRAS, 416, 689
Kipping, D. M. 2014, MNRAS, 444, 2263
Kipping, D. M., Bastien, F. A., Stassun, K. G., et al. 2014, ApJL, 785, L32
Kley, W., & Nelson, R. P. 2012, ARA&A, 50, 211
Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al. 2011, Natur, 470, 53
Lucy, L. B. 2013, A&A, 551, A47
Mandel, K., & Agol, E. 2002, ApJL, 580, L171
Matsumura, S., Takeda, G., & Rasio, F. A. 2008, ApJL, 686, L29
Moorehead, A. V., Ford, E. B., Morehead, R. C., et al. 2011, ApJS, 197, 1
Morton, T. D., & Johnson, J. A. 2011, ApJ, 729, 138
Morton, T. D., & Winn, J. N. 2014, ApJ, 796, 47
Naoz, S., Farr, W. M., Lithwick, Y., Rasio, F. A., & Teysandier, J. 2011, Natur, 473, 187
Naoz, S., Farr, W. M., & Rasio, F. A. 2012, ApJL, 754, L36
Parviainen, H., Deeg, H. J., & Belmonte, J. A. 2013, A&A, 550, A67
Payne, M. J., Deck, K. M., Holman, M. J., & Perets, H. B. 2013, ApJL, 775, L44
Plavchan, P., Bilinski, C., & Currie, T. 2014, PASP, 126, 34
Plummer, M. 2003, in Proc. 3rd Int. Workshop on Distributed Statistical Computing (DSC 2003), ISSN 1609-395X
Ragozzine, D., & Wolf, A. S. 2009, ApJ, 698, 1778
Rasio, F. A., & Ford, E. B. 1996, Sci, 274, 954
Rauer, H., et al. 2014, ExA
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, Proc. SPIE, 9143, 20
Rogers, L. A. 2015, ApJ, 801, 41
Steffen, J. H., Fabrycky, D. C., Agol, E., et al. 2013, MNRAS, 428, 1077
Steffen, J. H., Fabrycky, D. C., Ford, E. B., et al. 2012b, MNRAS, 421, 2342
Steffen, J. H., Ragozzine, D., Fabrycky, D. C., et al. 2012a, PNAS, 109, 7982
Sterne, T. E. 1940, PNAS, 26, 36
Tenenbaum, P., Jenkins, J. M., Seader, S., et al. 2014, ApJS, 211, 6
Torres, G., Winn, J. N., & Holman, M. J. 2009, in IAU Symp. 253, ed. F. Pont, D. Sasselov & M. J. Holman, 482
Valencicchi, F., & Rasio, F. A. 2014, ApJ, 786, 102
Van Eylen, V., & Albrecht, S. 2015, ApJ, 808, 126
Wang, J., & Ford, E. B. 2011, MNRAS, 418, 1822
Winn, J. N. 2010, Exoplanet Transits and Occultations, ed. S. Seager
Wolfgang, A., & Lopez, E. 2015, ApJ, 806, 183
Wu, Y., & Lithwick, Y. 2013, ApJ, 772, 74