



# Implicit Biases in Transit Models Using Stellar Pseudo Density

Gregory J. Gilbert<sup>1</sup>, Mason G. MacDougall<sup>1</sup>, and Erik A. Petigura<sup>1</sup>Department of Physics and Astronomy, University of California, Los Angeles, USA; [gjgilbert@astro.ucla.edu](mailto:gjgilbert@astro.ucla.edu)

Received 2022 February 9; revised 2022 June 3; accepted 2022 July 5; published 2022 August 11

## Abstract

The transit technique is responsible for the majority of exoplanet discoveries to date. Characterizing these planets involves careful modeling of their transit profiles. A common technique involves expressing the transit duration using a density-like parameter,  $\tilde{\rho}$ , often called the “circular density.” Most notably, the Kepler project—the largest analysis of transit light curves to date—adopted a linear prior on  $\tilde{\rho}$ . Here, we show that such a prior biases measurements of impact parameter,  $b$ , due to the nonlinear relationship between  $\tilde{\rho}$  and transit duration. This bias slightly favors low values ( $b \lesssim 0.3$ ) and strongly disfavors high values ( $b \gtrsim 0.7$ ) unless the transit signal-to-noise ratio is sufficient to provide an independent constraint on  $b$ , a criterion that is not satisfied for the majority of Kepler planets. Planet-to-star radius ratio,  $r$ , is also biased due to  $r - b$  covariance. Consequently, the median Kepler DR25 target suffers a 1.6% systematic underestimate of  $r$ . We present a techniques for correcting these biases and for avoiding them in the first place.

*Unified Astronomy Thesaurus concepts:* Exoplanets (498); Bayesian statistics (1900); Astronomy data modelling (1859); Light curves (918); Transit photometry (1709)

## 1. Introduction

In the two decades since the discovery of the first transiting hot Jupiter (Charbonneau et al. 2000; Henry et al. 2000), the transit technique has grown to be the most prolific exoplanet detection method to date, accounting for 77% of the current census. Contemporary work continues to rely heavily on the transit technique. To wit, several transit-focused NASA and ESA missions are either already on-sky (TESS; Ricker et al. 2015) or slated for launch in the near future (PLATO; Rauer et al. 2014), and next-generation radial-velocity spectrographs have been designed for follow-up characterization of known transiting planets (e.g., KPF; Gibson et al. 2016; MAROON-X; Seifahrt et al. 2018). The transit technique will remain indispensable for exoplanet astronomy for decades to come.

Accurate modeling of the transit light curve is a critical step for characterizing transiting planets. At the most basic level, transit modeling involves computing the time-dependent flux  $F(t)$  of a star obscured by a transiting planet relative to the unobscured flux  $F_0$ . If one assumes a spherical planet and star, this computation depends strictly on the planet-to-star size ratio  $r$ , the (time-dependent) center-to-center sky-projected planet-to-star separation  $z$  (measured in units of  $R_*$ ), and the radial dependence of the stellar limb-darkening profile  $\{u\}$ . Early analyses computed  $F(z; r, \{u\})$  via numerical integration, but today the most widely used method is the Mandel & Agol (2002) model, which expresses the transit light curve via an analytic solution to  $F(z; r, \{u\})$  for several limb-darkening profiles that can be described by a small set of limb-darkening parameters.

In order to model time-series photometry, one must convert  $F(z; r, \{u\})$  into  $F(t; r, \{u\})$ . While  $z$  is the only parameter that varies with time, one may choose how to specify the function that maps  $t \rightarrow z$ . If one assumes strict periodicity of transits and a constant projected velocity during transit, then in the limit

$r \rightarrow 0$ ,  $z(t)$  may be specified completely by an orbital period,  $P$ , a transit midpoint,  $t_0$ , an impact parameter,  $b$ , and first-to-fourth contact transit duration,  $T_{14}$ .<sup>1</sup> This parameterization— $F(t; P, t_0, r, b, T_{14})$ —is convenient and is closely linked to the transit geometry.

An alternative approach is to specify  $T_{14}$  from a combination of scaled separation  $a/R_*$ , orbital eccentricity  $e$ , argument of periastron  $\omega$ , and projected inclination  $\cos i$ , following Winn (2010) as

$$T_{14} \simeq \frac{P}{\pi} \sin^{-1} \left( \frac{R_*}{a} \frac{\sqrt{(1+r)^2 - b^2}}{\sin i} \right) \left( \frac{\sqrt{1-e^2}}{1+e \sin \omega} \right) \quad (1)$$

$$b = \frac{a \cos i}{R_*} \left( \frac{1-e^2}{1+e \sin \omega} \right). \quad (2)$$

Now the light curve is specified by the function  $F(t; P, t_0, r, a/R_*, b, e, \omega)$ , which is similar to the parameterization used by the EXOFAST suite (Eastman et al. 2013; Eastman 2017).<sup>2</sup> A related approach is to replace  $a/R_*$  with stellar density by employing Kepler’s third law. Thus, the light curve may also be parameterized by  $F(t; P, t_0, r, \rho_*, \cos i, e, \omega)$ .

These two eccentricity-explicit parameterizations have the advantage that the light curve has been specified completely by properties of the star, planet, and planetary orbit; the disadvantage is that five parameters have been replaced by seven, and thus significant degeneracies between  $\{a/R_*, e, \omega\}$  or  $\{\rho_*, e, \omega\}$  are inevitable. These degeneracies lead to inefficiencies with light-curve fitting and posterior sampling.

<sup>1</sup> Several alternative transit durations may be substituted for  $T_{14}$ : (1) the second-to-third contact duration,  $T_{23}$ , (2) the center-to-center contact duration,  $T_{cc}$ , also called the 1.5-to-3.5 contact duration, or (3) the full-width-half-max duration,  $T_{FWHM}$ , which may be defined in relation to the transit depth. While each has its merits (see Kipping 2010a for discussion), we adopt  $T_{14}$  throughout this work because it is the transit duration which is most readily defined for all grazing and non-grazing transit geometries.

<sup>2</sup> In practice, EXOFAST uses  $\log(a/R_*)$  and expresses  $b$  as  $\cos i$ ;  $\{e, \omega\}$  is usually specified as  $\{\sqrt{e} \sin \omega, \sqrt{e} \cos \omega\}$  in order to establish uniform priors on  $e$  and  $\omega$  and to avoid a boundary issue at  $e = 0$ .



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

A common shortcut is to fit the light curve assuming that  $e=0$  even though the orbit may, in fact, be eccentric. This assumption reduces the number of free parameters back to five, but  $\rho_*$  can no longer be thought of as a stellar density. Rather, it is a stand-in for duration which merely has *units* of density, defined by Seager & Mallén-Ornelas (2003) as

$$\tilde{\rho} \equiv \left( \frac{4\pi^2}{P^2 G} \right) \left( \frac{(1+r)^2 - b^2(1 - \sin^2[\pi T_{14}/P])}{\sin^2[\pi T_{14}/P]} \right)^{3/2} \quad (3)$$

where  $G$  is Newton’s gravitational constant. This quantity  $\tilde{\rho}$  is sometimes referred to as the “mean stellar density,” the “circular density,” or the “observed density,” but we prefer to call it the “pseudo density” because (1) the other names are confusing, and (2)  $\tilde{\rho}$  matches the true stellar density only when numerous assumptions are met (see Kipping 2014).

Because the prior expectation for  $\tilde{\rho}$  is a complicated function of  $\rho_*$ ,  $b$ ,  $e$ , and  $\omega$ , naively placing a flat prior on  $\tilde{\rho}$  and adopting it as a fitting parameter induces undesired biases on  $T_{14}$  and  $b$ .

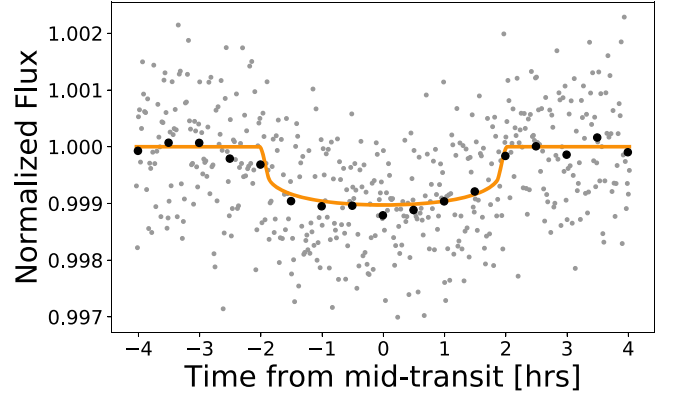
To date,  $\tilde{\rho}$  has enjoyed widespread use in the exoplanet literature. For example, the Kepler project (Borucki et al. 2010); the largest analysis of transit light curves to date) fit their light curves with the  $F(t; P, t_0, r, \tilde{\rho}, b)$  parameterization (Rowe et al. 2014, 2015; Mullally et al. 2015; Coughlin et al. 2016; Thompson et al. 2018). We discuss the effects of that choice in Section 5. More broadly, this paper investigates the implicit biases on impact parameter and other light-curve parameters that result from the use of  $\tilde{\rho}$ .

Throughout this work, we assume that all transit signals under investigation have been thoroughly vetted such that the detected signal is known to be a real transit at high confidence. The methods employed in this work are thus appropriate for parameter estimation but not for transit detection or vetting.

This paper is organized as follows. In Section 2 we empirically demonstrate the origin of the  $\tilde{\rho}$  bias by fitting a transit light-curve model to simulated photometry using the Kepler project parameterization; we then demonstrate that our preferred parameterization does not suffer from this bias. In Section 3 we present a numerical experiment which isolates the effects of various model assumptions on posterior inferences. In Section 4 we analytically derive the Jacobian of the coordinate transformation  $T_{14} \rightarrow \tilde{\rho}$  which explains the origin of the empirical bias. In Section 5 we show that the  $\tilde{\rho}$  bias has affected most posterior inferences of  $b$  and  $r$  derived from Kepler data. In Section 6 we summarize our conclusions and discuss other biases which arise from using related parameterizations such as  $a/R_*$ .

## 2. Understanding Parameter Biases with Fits to Synthetic Photometry

To illustrate the  $\tilde{\rho}$  bias, we simulated photometric observations of a warm mini-Neptune ( $P = 15$  days,  $r_p = 3.3 R_\oplus$ ) on a circular orbit around a Sun-like star, transiting at impact parameter  $b = 0.5$ . We simulated data with a 30 minute observing cadence (matching Kepler’s long-cadence observing mode) within  $\pm T$  from the transit center. All photometric data were oversampled by a factor of 7 and integrated using Simpson’s rule to account for the effects of finite integration time (Kipping 2010b). The white noise level was tuned to produce signal-to-noise ratio (S/N) = 16, which is slightly lower than the median Kepler value and results in a posterior model with  $\sigma_r/r \approx 0.10$  and  $\sigma_T/T \approx 0.05$ , where  $\sigma_r/r$



**Figure 1.** Simulated photometry for a mini-Neptune on a circular 15 day orbit around a Sun-like star, transiting at  $b = 0.5$ . The orange line indicates the ground-truth transit model. Gray points show simulated observations at a one minute observing cadence; black circles are binned to 30 minutes. The white noise level was set to S/N = 16, close to the Kepler median. See Table 1 for ground-truth simulation parameters.

**Table 1**  
Ground-truth Simulation Parameters<sup>a</sup>

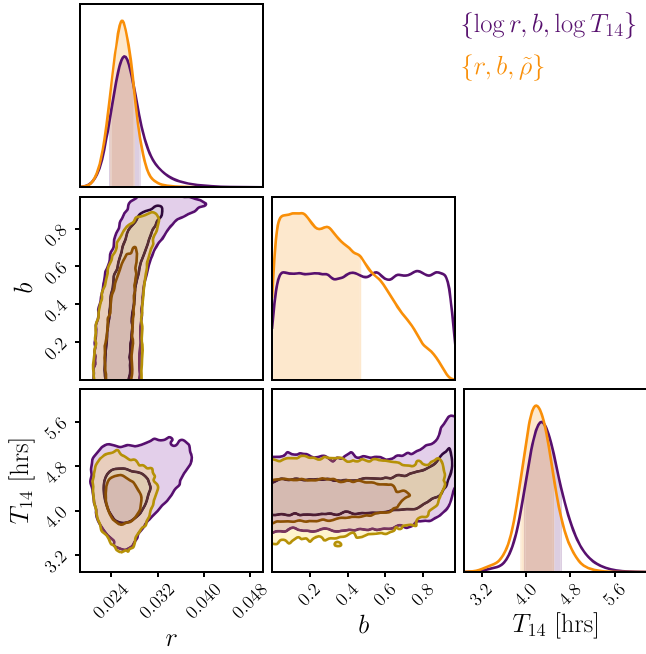
Parameter	Value
$M_*/[M_\odot]$	1.0
$R_*/[R_\odot]$	1.0
$u_1, u_2$	0.40, 0.25
$P$ [days]	15.0
$r$	0.03
$b$	0.5
$T_{14}$ [hrs]	3.29
S/N	16

**Note.**

<sup>a</sup> Simulated photometry is shown in Figure 1.

corresponds to the fractional posterior measurement, and similar for  $T$ . We chose these values in order to produce a transit that is similar to those found by Kepler. Ground-truth simulation parameters are listed in Table 1, and simulated photometry is shown in Figure 1.

The transit model was specified using a standard pseudo-density parameterization:  $\{P, t_0, r, b, \tilde{\rho}\}$ . In order to minimize confounding factors, we held  $P$  and  $t_0$  fixed at their injected values; we also held the mean out-of-transit flux,  $F_0$ , and photometric white noise level,  $\sigma_{\text{phot}}^2$ , fixed to their true values, which is equivalent to assuming the raw photometry has been accurately prewhitened. For the remaining transit parameters, we adopted broad weakly informative priors with permissive bounds (see Table 2 for details), for a total of three free parameters per model:  $\{r, b, \tilde{\rho}\}$  (the “ $\tilde{\rho}$  basis”), or  $\{\log r, b, \log T_{14}\}$  (the “log  $T$  basis”). We chose the later basis because  $T_{14}$  is typically well constrained by the data and furthermore may be assigned priors in a sensible fashion; sampling in  $\log r$  and  $\log T_{14}$  is equivalent to placing log-uniform priors on  $r$  and  $T_{14}$ , which facilitates exploration of posterior values over different orders of magnitude. We modeled a circular transit in all cases and held stellar mass, radius, and limb darkening to their true values during the fit; there is no loss of generality in this approach because as long as we ignore minuscule ingress/egress asymmetry that exists for eccentric transits (Barnes 2007), there is no difference between a circular and eccentric transit. In order to avoid complications which arise when modeling grazing transits, we restricted impact



**Figure 2.** Posterior corner plots when sampling in the  $\tilde{\rho}$  basis (orange) vs. the  $\log T_{14}$  basis (purple). The bias on impact parameter,  $b$ , is apparent when sampling with  $\tilde{\rho}$  but is resolved when sampling in  $T_{14}$ .

**Table 2**  
Priors on Model Parameters for Simulated Light Curve

Parameter	Value	Parameter	Value
$r$	$\mathcal{U}(0.01, 0.1)$	$\log r$	$\mathcal{U}(-2, -1)$
$b$	$\mathcal{U}(0, 1 - r)$	$b$	$\mathcal{U}(0, 1 - r)$
$\tilde{\rho}/\rho_{\odot}$	$\mathcal{U}(0.1, 10)$	$\log[T_{14}/\text{hr}]$	$\mathcal{U}(1, 10)$

parameters to  $b < 1 - r$ .<sup>3</sup> To confirm that this restriction is permissible, we explored the parameter space near the limb of the star following the methodology of Gilbert (2022) and verified that the simulated transit is inconsistent with a grazing geometry.

We drew samples from the posterior using Hamiltonian Monte Carlo (Neal 2011) and the No U-Turn Sampler (Hoffman & Gelman 2011). Each model iteration consisted of two chains run for 5000 tuning steps and 20,000 draws, producing an effective number of samples greater than 11,000 for all parameters for each of the two parameterizations.

Posterior corner plots for the quantities of interest are shown in Figure 2. The most notable difference is in the 1D marginalized distribution of impact parameter. When sampling using the  $\tilde{\rho}$  basis, the posterior is biased toward low  $b$ ; as a point of reference, 74% of the probability mass is below  $b = 0.5$ , the injected value. When sampling using the  $\log T$  basis, however, the distribution of  $b$  is nearly uniform over the allowed range, reflecting the fact that for a low signal-to-noise transit the impact parameter is largely unconstrained. Our results did not substantially change when simulating a one minute observing cadence (matching Kepler’s short cadence mode), indicating that the  $\tilde{\rho}$  bias arises from the model

<sup>3</sup> A common approach (which we did not adopt) is to draw samples uniformly from the  $r - b$  plane using triangular sampling (Espinoza 2018). However, naive application of this method induces a marginal prior on  $r$ , so caution must be taken to ensure that priors are established as intended.

parameterization and is not an artifact of data binning. We also repeated the analysis using  $r = 0.1$  and  $r = 0.01$  and found that the results did not change.

Clearly, the results are inconsistent between models—which contain identical underlying physics and differ only in their parameter bases—so at least one of the two models has produced biased inference. In the sections that follow, we present both a numerical argument (Section 3) and an analytic argument (Section 4), which demonstrate that the  $\log T$  basis has produced the desired result.

### 3. Numerical Sampling Experiment

We will now demonstrate that the bias on  $b$  seen in the previous section arises solely from the model parameterization and not from vagaries of the Markov chain Monte Carlo (MCMC) sampling algorithm or peculiarities of the noise realization in the photometry.

To do so, we performed a numerical experiment that approximated the light-curve modeling procedure from Section 2 by drawing samples directly from the prior distributions and then applying an a posteriori importance weighting designed to mimic the constraints imposed by the photometry. When determining these importance weights, we employed a Gaussian-likelihood function and approximated the (covariant) parameter constraints from Section 2 as independent univariate Gaussians. The key advantage of this method is that we no longer needed to directly fit the photometry, thereby eliminating potential confounding factors introduced by the photometry and the sampler.

#### 3.1. Experimental Setup

We adopted the same fiducial star-planet system as Section 2, placing a  $3.3R_{\oplus}$  mini-Neptune on a circular 15 day orbit around a solar twin. We fixed the ephemeris  $\{P, t_0\}$  throughout and placed uniform interval priors on all other parameters  $\{r, b, \tilde{\rho}, \log T_{14}\}$  as before (see Table 2), with the small modification that we now allow  $b$  to range over all detectable values, i.e.,  $b \sim \mathcal{U}(0, 1 - r)$ ; this modification is acceptable because our sampling procedure (see below) avoids the usual issues that arise when fitting grazing transits (see Gilbert 2022).

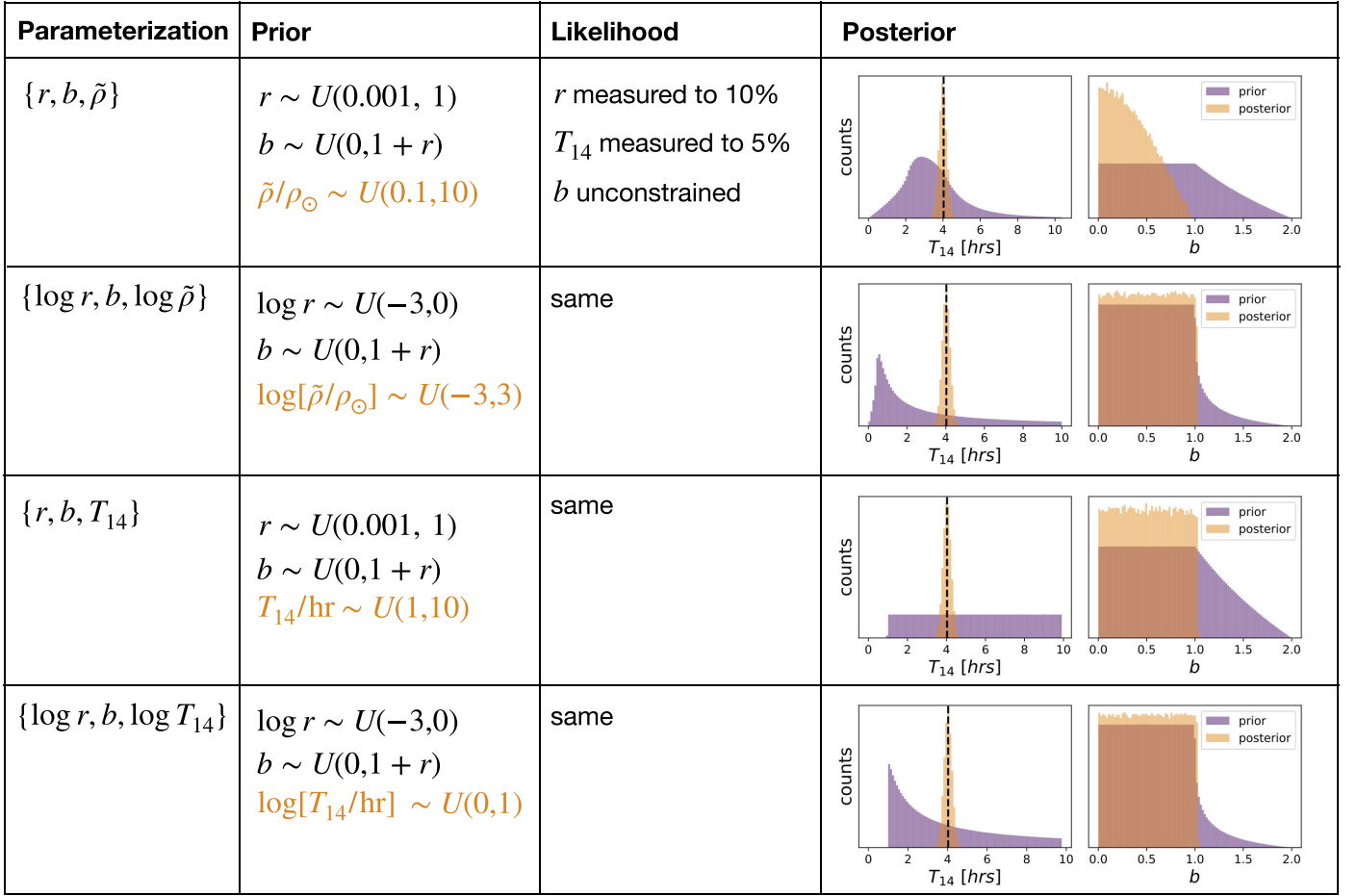
For the first iteration of the experiment we adopted the  $\tilde{\rho}$  basis  $\{r, b, \tilde{\rho}\}$  and drew random samples directly from the prior distributions. We next calculated transit duration using

$$T_{14} = \frac{PR_{\star}}{\pi a} ((1 + r)^2 - b^2)^{1/2} \quad (4)$$

for each sample. Here, we have approximated Equation (1) by using the small angle approximation  $\sin^{-1} \phi \approx \phi$  and  $i \approx \pi/2 \rightarrow \sin i \approx 1$ . The scaled separation can be calculated from Kepler’s Third Law as  $a/R_{\star} = [(GP^2\rho)/(3\pi)]^{1/3}$ .

For subsequent iterations of the experiment, we modified the procedure to use three alternative parameter bases: (1)  $\{\log r, b, \log \tilde{\rho}\}$ , (2)  $\{r, b, T_{14}\}$ , and (3)  $\{\log r, b, \log T_{14}\}$ . We chose these parameterizations in order to explore the effects of uniform versus log-uniform priors in addition to the effect of substituting  $\tilde{\rho} \rightarrow T_{14}$ . We followed the same sampling procedure as before, except when drawing samples of  $T_{14}$  or  $\log T_{14}$  we calculated  $\tilde{\rho}$  following Equation (3).

Mimicking the simulated light curve in Section 2, we assumed that we could constrain  $r$  to 10% accuracy and  $T_{14}$  to



**Figure 3.** Results of the numerical sampling experiment described in Section 3. Each row corresponds to the prior, likelihood, and posterior for a given model parameterization. For visual clarity, the height of the  $T_{14}$  posterior has been reduced by a factor of 3 on all plots. The difference in the prior distribution on  $b$  for rows 1 & 3 compared to rows 2 & 4 stems from the use of  $r$  vs.  $\log r$ , respectively. Sampling with a uniform prior on  $\tilde{\rho}$  (top row) produces a nonuniform prior on  $T_{14}$  and a biased posterior for  $b$ . In contrast, sampling in any of the other parameter bases produces a posterior estimate of  $b$ , which matches the prior, except in cases where constraints on  $r$  would produce a nontransiting orbit.

5% accuracy, with independent Gaussian precision from the photometry (i.e.,  $\sigma_r/r = 0.1$ ,  $\sigma_T/T = 0.05$ ). We further assumed that the impact parameter would be entirely unconstrained by the data. These uncertainties are representative of typical values, but we have removed the covariance and forced them to be Gaussian (or unconstrained), which eases interpretation.

We imposed our assumed measurement uncertainties on  $r$  and  $T_{14}$  by calculating the log likelihood of each  $i$ th sample

$$\log \mathcal{L}_i = -\frac{1}{2} \left( \frac{T_i - T_{\text{true}}}{\sigma_T} \right)^2 - \frac{1}{2} \left( \frac{r_i - r_{\text{true}}}{\sigma_r} \right)^2, \quad (5)$$

which assumes a Gaussian-likelihood function. We then weighted each sample by

$$w_i = \frac{\mathcal{L}_i}{\sum_i \mathcal{L}_i} \quad (6)$$

to produce our synthetic posterior distributions.

### 3.2. Bias on Impact Parameter

The results of our numerical experiment are summarized in Figure 3. As expected, when parameterizing the model as  $\{r, b, \tilde{\rho}\}$  with uniform priors, we obtain biased results that are qualitatively similar to those produced in Section 2 (i.e., by

fitting the photometry directly). Notably, sampling in  $\tilde{\rho}$  produces a strong prior on  $T_{14}$  (purple) that is not physically motivated. Because  $T_{14}$  is constrained to 5%, the data overwhelm the prior and the  $T_{14}$  posterior distribution (orange) is only slightly biased. The posterior on impact parameter, however, is clearly different from the prior even though our model included no information about impact parameter. Because we have (by construction) placed no measurement constraint on  $b$ , the posterior distribution should match the prior. In reality however, the posterior is tilted toward  $b = 0$ , giving the illusion of a (modestly) constrained posterior.

The  $\tilde{\rho} - b$  bias is resolved by using any of the alternative parameterizations which substitute  $\log \tilde{\rho}$ ,  $T_{14}$ , or  $\log T_{14}$  for  $\tilde{\rho}$ . Although using the substitution  $\tilde{\rho} \rightarrow \log \tilde{\rho}$  may seem at first glance to be the simplest choice (requiring little change from existing practices), we argue that using either of the duration-based parameterizations is preferable for two reasons. First, the results are insensitive to the exact choice of (reasonable) prior placed on  $T_{14}$ , whereas they are highly sensitive to the prior placed on  $\tilde{\rho}$ ; insensitivity to priors is in general a desirable feature of robust inference. Second, setting prior interval bounds on  $\tilde{\rho}$  is a nonintuitive task, requiring careful consideration of the true stellar density and orbital elements. In contrast, principled priors may be placed on the transit duration quite simply following inspection of the transit light curve. In fact, setting bounds on  $T_{14}$  is so



straightforward that it could even be done algorithmically following the output of a box-least-squares transit search (Kovács et al. 2002). The bottom line is that given the choice between options which produce equivalent results, we prefer the simpler of the two.

In summary, because we have decoupled the posteriors from complicating factors (e.g., parameter covariances, sampler inefficiencies, etc.), we conclude that the differences between posterior distributions obtained under the  $\tilde{\rho}$  basis versus the  $\log T_{14}$  basis arise solely due the parameterization. Furthermore, we conclude that the  $\tilde{\rho}$  basis (with a uniform prior) induces a bias on  $b$ , whereas the other options we have presented produce unbiased estimates.

#### 4. Mathematical Origin of the Bias

In the previous sections, we illustrated the biases on  $b$  that result from uniform and log-uniform priors on  $\tilde{\rho}$  by exploring synthetic photometry fits and simple numerical experiments. In this section, we investigate the mathematical origins of this bias.

The transit parameter covariance matrix was previously derived by Carter et al. (2008), but where their treatment prioritized analytic interpretability (with a small sacrifice to accuracy), our treatment prioritizes accuracy (with a small sacrifice to interpretability). Most importantly, the covariance matrix derived by Carter et al. (2008) are least accurate as  $b \rightarrow 1$  and in the presence of nonnegligible limb darkening, which are precisely the conditions under which the  $\tilde{\rho}$  bias we are investigating become most important. Thus, our work complements rather than supplants Carter et al. (2008).

When modeling light curves, our main goal is to derive the posterior probability density function,  $p(\mathbf{x})$ , i.e., the probability that a set of planet properties  $\mathbf{x}$  resides in an infinitesimal volume element spanning  $\mathbf{x}$  to  $\mathbf{x} + d\mathbf{x}$ . However, this probability is not invariant under changes in parameterization. Specifically, for our problem,  $p(T_{14})/dT_{14} \neq p(\tilde{\rho})/d\tilde{\rho}$ . To convert  $p(T_{14})$  to  $p(\tilde{\rho})$ , one must account for the change in infinitesimal volume element resulting from the  $T_{14} \rightarrow \tilde{\rho}$  transformation, i.e., the Jacobian

$$J = \frac{d\tilde{\rho}}{dT_{14}} = -\frac{12\pi^3}{P^3G}((1+r)^2 - b^2)^{3/2} \left( \frac{\pi T_{14}}{P} \right)^{-4}, \quad (7)$$

which we derive in the Appendix. The Jacobian of the transformation  $T_{14} \rightarrow \log \tilde{\rho}$  is simply

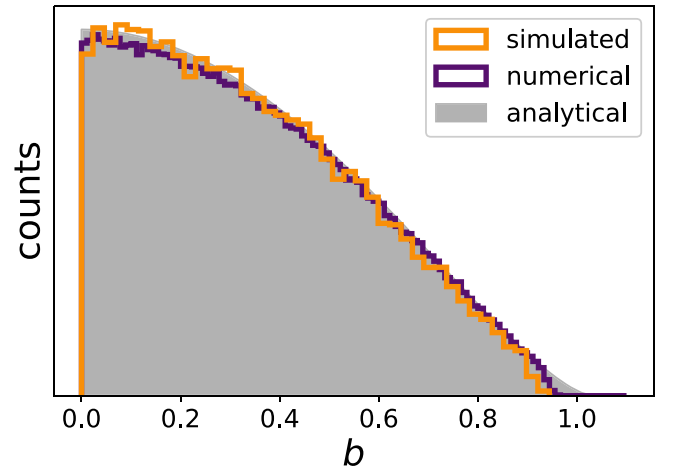
$$J' = \frac{d \log \tilde{\rho}}{dT_{14}} = -\frac{3}{T_{14}}, \quad (8)$$

which is independent of  $b$ , explaining why using  $\log \tilde{\rho}$  in place of  $\tilde{\rho}$  produces unbiased posteriors.

In Figure 4, we show the analytic Jacobian in Equation (7) alongside the simulated posterior samples of  $b$  obtained in Section 2 and the numerical results obtained in Section 3. It is evident from inspection that the distributions are in close agreement. We conclude that the nonuniform distribution of  $b$  arises from the combination of parameterization and (incorrect) prior, rather than from any real constraint imposed by the data.

#### 5. Biased Kepler Planet Properties

We have shown that adopting a linear  $\tilde{\rho}$  prior results in a biased impact parameter. The Kepler project (Borucki et al. 2010; Rowe et al. 2014, 2015; Mullally et al. 2015; Coughlin et al. 2016; Thompson et al. 2018) used such a

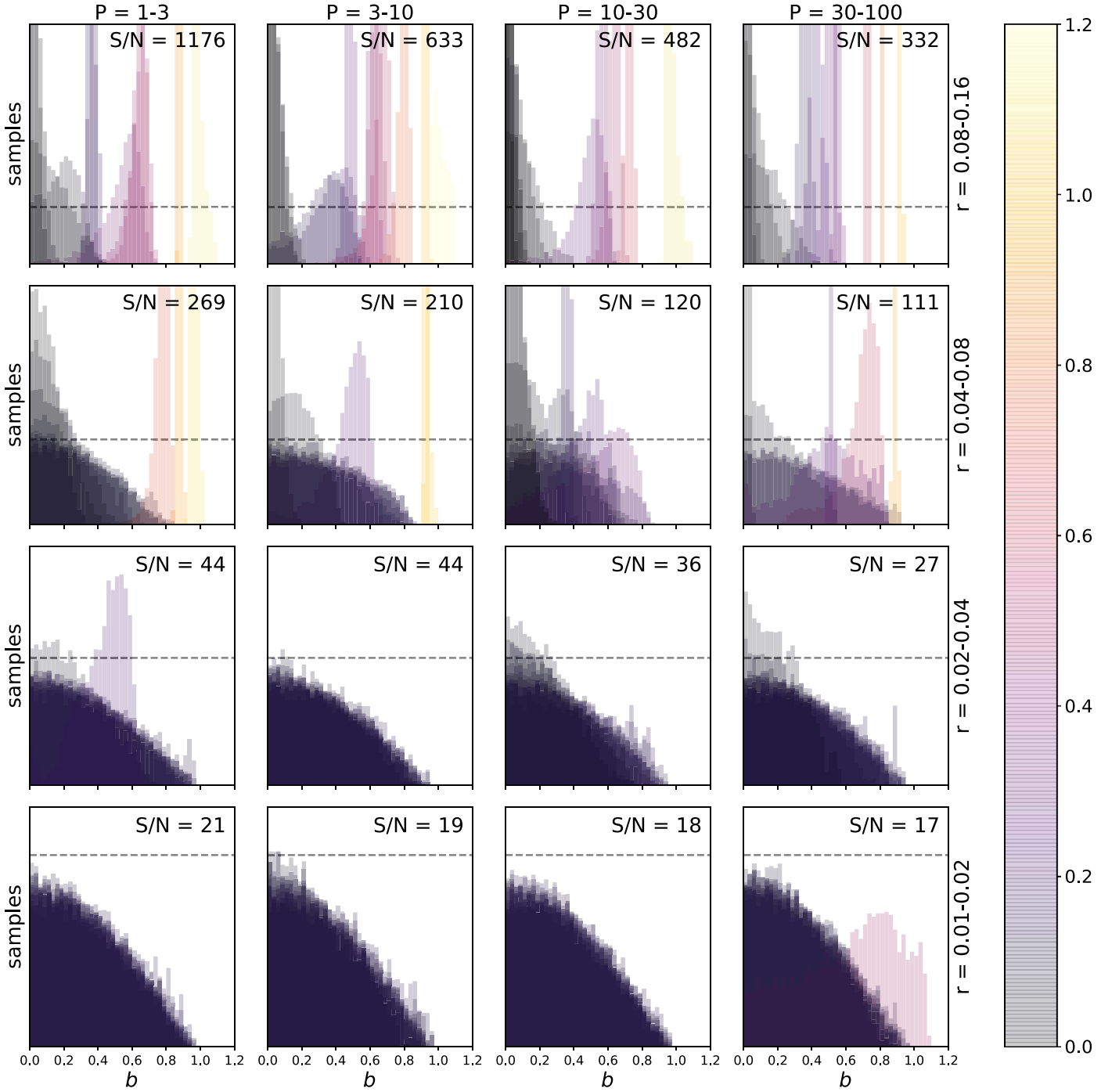


**Figure 4.** Posterior samples of  $b$  from the simulated transit fit (orange histogram, Section 2) and the numerical experiment (purple histogram, Section 3) are nearly perfectly matched by the expected bias from the analytically derived Jacobian (gray shaded region, Section 4).

parameterization (J. Rowe 2022, private communication). Therefore, we expect biased  $b$  in all cases except those where  $b$  is strongly constrained by the light curve itself. Because most Kepler planet candidates exhibit modest transit signal-to-noise (median  $S/N = 22.4$ ), the characteristic “hill” shape we have seen for biased posterior  $b$  distributions in the previous three sections is also present in the posterior distributions of nearly every Kepler planet candidate from DR25 (Thompson et al. 2018). Figure 5 illustrates the presence of the  $b$  bias over a grid of orbital periods and radii. Only the largest (and therefore highest  $S/N$ ) planets consistently exhibit meaningful constraints on  $b$ .

Due to signal-to-noise bias which disfavors the detection of high- $b$  transits (Kipping & Sandford 2016), the prior expectation on impact parameter is not exactly flat, and so the posteriors exhibited in Kepler data will not exactly match the idealized distribution we derived in Sections 2–4. However, most Kepler detections have  $S/N > 10$  and fall in the flat part of the detection completeness curve (Christiansen et al. 2020). Thus, the appropriate prior for the vast majority of Kepler planets should be nearly flat in  $b$ , with a fall off at the value of  $b$  that reduces  $S/N$  to  $\sim 10$ .

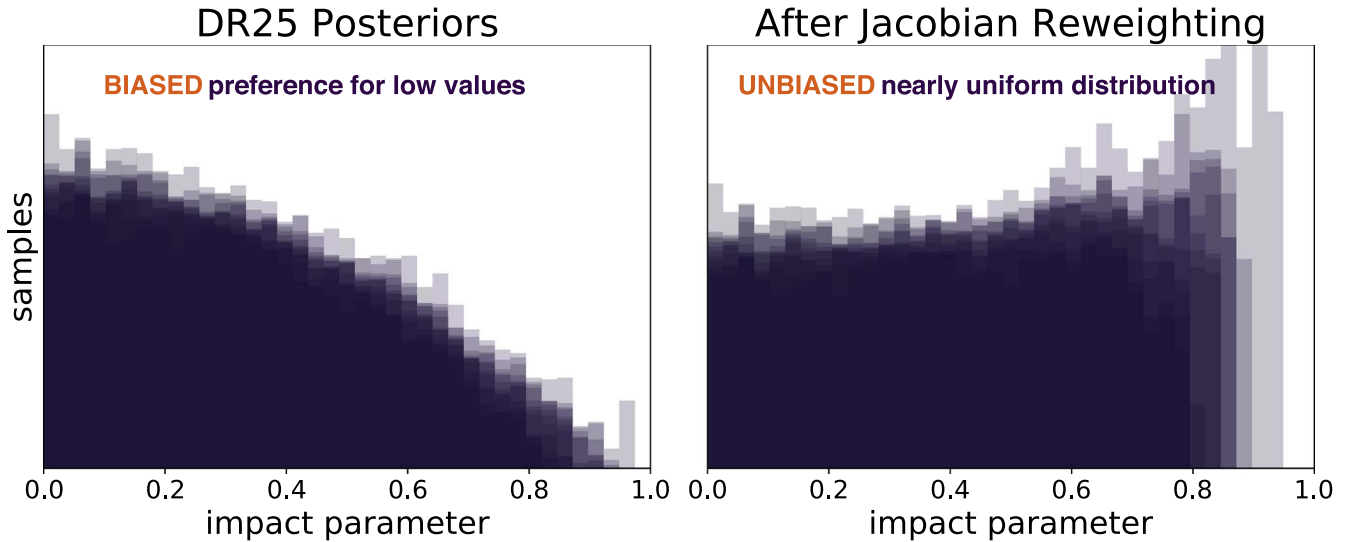
Detection biases notwithstanding, the  $\tilde{\rho}$  bias is easily understood and corrected. Because the relationship between  $\tilde{\rho}$ ,  $b$ ,  $r$ , and  $T_{14}$  is known analytically (Seager & Mallén-Ornelas 2003), one needs only to apply the appropriate Jacobian weighting in order to transform an unintended prior on  $\tilde{\rho}$  into the desired prior on  $b$  or  $T_{14}$  (or any other basis parameter derivable from these quantities). Unbiased parameter estimates can then be recovered from existing (biased) posterior chains by implementing an importance sampling scheme that accounts for this coordinate transformation, provided the chains are not too sparsely sampled in their low probability regions. Specifically, one can sample from a distribution  $p_1(\mathbf{x})$  by reweighting samples from a different distribution  $p_2(\mathbf{x})$ . An example of this reweighting scheme as applied to a selection of DR25 targets is shown in Figure 6. A caveat is there is increased sampling error since  $p_2(\mathbf{x})$  is a different distribution and the samples are not optimally distributed in  $p_1(\mathbf{x})$ . In essence there are smaller number of “effective samples” after reweighting. Care must therefore be taken to ensure that Jacobian-corrected posteriors are reliable, and the reweighting scheme we have outlined here should not be applied blindly.



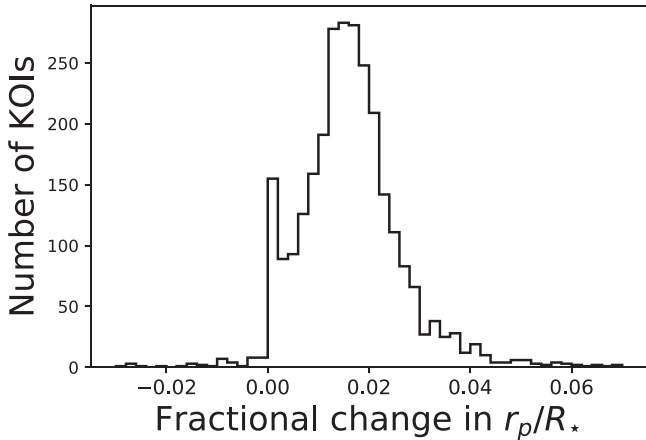
**Figure 5.** Posterior distributions of impact parameter for a random selection of KOIs, organized in logarithmic bins on a  $P-r$  grid. Data shown are the posterior MCMC chains from Kepler Data Release 25 (Thompson et al. 2018), described in detail in Rowe et al. (2014) and downloaded from the NASA Exoplanet Archive (Akeson et al. 2013). Each posterior distribution is plotted with 20% opacity so that dark regions indicate where many distributions overlap; colors correspond to the median  $b$  value for a given KOI. For visual clarity, a maximum of twelve KOIs are plotted per panel. The horizontal axis of each panel ranges over  $b = (0, 1.2)$ ; the vertical range of each row is different, but the dashed line indicates the same distribution height. The median S/N in each 2D bin is printed in the upper right-hand corner of the grid squares. It is clear from inspection that most of the objects (excluding the largest, highest-S/N objects) show qualitatively similar posterior distributions of  $b$ . The similarity is particularly striking for small (low-S/N) objects.

Because  $b$  is covariant with  $r$  (interacting via non-zero limb darkening), any bias on  $b$  translates to a bias on  $r$ . For measurements in the final Kepler data release, DR25, we find this covariance has produced a 1.6% median systematic underestimate of  $r$  (Figure 7), extending as high as  $\sim 6\%$  for some targets. This offset is comparable to the fractional uncertainty on

$R_*$  (Gaia Collaboration et al. 2018; Berger et al. 2018) and so makes up a sizeable portion of the error budget for Kepler planetary radii. While a few percent difference in planetary radius for a *single* planet may be subsignificant, a systematic bias of a few percent on *all* planetary radii will significantly impact our interpretation of population demographics—for example, the



**Figure 6.** Posterior samples of  $b$  from a representative selection of DR25 targets before and after reweighting by the Jacobian to correct for biases induced by sampling in  $\tilde{\rho}$ . All targets have  $0.02 < r < 0.04$  and  $10 < P < 30$  days. Left panel: raw DR25 posteriors chains show a clear (biased) preference for low values of  $b$ . Right panel: after reweighting, the (unbiased) distribution is nearly flat. To minimize spurious peaks and sampling noise in low probability regions, the lowest density 1% of samples have been excluded from our reweighting scheme. The slight increase in probability density near  $b \approx 1$  in the reweighted posteriors reflects the presence of residual importance sampling noise rather than a real feature of the data. Because there is significant sampling noise (due to the large implied posterior mass in regions with few samples), our preferred method for ameliorating the pseudo-density bias is to refit the photometry.



**Figure 7.** Fractional change in median planet-to-star radius ratio for all planet candidates after correcting posterior chains from DR25 (Thompson et al. 2018) using the Jacobian reweighting scheme described in Section 5.  $5\sigma$  outliers have been iteratively clipped in order to eliminate spurious values that are expected to arise due to insufficient sampling of low probability regions. There is a spike at  $\delta_r = 0$ , indicating that some subset of targets were accurately measured, but the majority of targets are distributed around  $\delta_r/r = 1.6\%$ .

precise characteristics of the radius valley (Fulton et al. 2017)—thereby altering our understanding of the processes by which planets form and evolve.

## 6. Summary and Conclusions

In this work, we explored the biases that result from using the popular stellar pseudo-density,  $\tilde{\rho}$ , as a parameter in light-curve fits. Adopting a linear prior on this parameter results in a biased distribution on impact parameter due to the Jacobian that arises from the nonlinear relationship between  $\tilde{\rho}$  and transit duration,  $T_{14}$ . Biased inferences on  $b$  lead to biased inferences on  $r$  due to covariances between the two parameters. We confirmed that these biases are present in Kepler modeling that used  $\tilde{\rho}$  as a fitting parameter, and we presented a method for debiasing the distributions.

Although the  $\tilde{\rho}$  bias may be resolved by using  $\log \tilde{\rho}$  in place of  $\tilde{\rho}$  (or, equivalently, placing log-uniform priors on  $\tilde{\rho}$ ), we prefer sampling in duration over  $\tilde{\rho}$  for esthetic and conceptual reasons. To avoid inducing biases, we recommend sampling directly in duration  $T_{14}$  or replacing  $T_{14}$  with the true stellar density and orbital eccentricity vector, i.e.,  $\{\rho_*, \sqrt{e} \sin \omega, \sqrt{e} \cos \omega\}$ .

This work focused on the biases induced from using  $\tilde{\rho}$  directly as a fitting parameter; similar biases may arise when using any related parameterization, for example  $a/R_*$ , which is a popular choice (e.g., Crossfield et al. 2015; David et al. 2016; Stassun et al. 2017). As with  $\tilde{\rho}$ , adopting a log-uniform prior rather than a linear prior on  $a/R_*$  avoids the unwanted bias. A log-uniform prior is a common choice, so most analyses which have used  $a/R_*$  as a fitting parameter are probably unaffected by the bias. However, one should always verify what priors were adopted when interpreting the results of any transit model.

We thank the anonymous referee for reviewing and providing comments which improved the quality of this manuscript. We are grateful to Jason Eastman, Dan Fabrycky, Dan Foreman-Mackey, Jason Rowe, Josh Winn, and Jon Zink for helpful conversations about this work.

G.J.G., M.G.M., and E.A.P. acknowledge support from NASA Astrophysics Data Analysis Program (ADAP) grant (80NSSC20K0457). E.A.P. acknowledges support from the Alfred P. Sloan Foundation. M.G.M. acknowledges support from the UCLA Cota-Robles Graduate Fellowship.

This study made use of data products from the Kepler mission hosted on the NASA Exoplanet Archive (2021). Some of the data were obtained from the Mikulski Archive for Space Telescopes (MAST) at the Space Telescope Science Institute. These data can be accessed via [10.26133/NEA5](https://exoplanetarchive.nasa.gov). This study also made use of computational resources provided by the University of California, Los Angeles and the California Planet Search.

*Facilities:* Kepler.

*Software:* astropy (Astropy Collaboration et al. 2018), exoplanet (Foreman-Mackey et al. 2021), numpy (Harris et al. 2020), PyMC

(Salvatier et al. 2016), scipy (Virtanen et al. 2020), starry (Luger et al. 2019).

### Appendix A Derivation of Jacobian for $T_{14} \rightarrow \tilde{\rho}$

In this section, we derive the Jacobian of the coordinate transformation  $T_{14} \rightarrow \tilde{\rho}$ . The pseudo density derived by Seager & Mallén-Ornelas (2003) is

$$\tilde{\rho} \equiv \left( \frac{4\pi^2}{P^2 G} \right) \left( \frac{(1+r)^2 - b^2(1 - \sin^2[\pi T/P])}{\sin^2[\pi T/P]} \right)^{3/2} \quad (\text{A1})$$

where all variables are defined as in previous sections. For notational clarity, we also define  $T \equiv T_{14}$  and make the simplifying assumption  $r \approx \sqrt{\Delta F}$ , where  $\Delta F$  is the fractional change in flux. Substituting terms

$$\begin{aligned} x &= 4\pi^2/(P^2 G) \\ y &= (1+r)^2 \\ z &= \sin^2[\pi T/P] \end{aligned} \quad (\text{A2})$$

yields

$$\tilde{\rho} = x \left( \frac{y - b^2(1-z)}{z} \right)^{3/2}. \quad (\text{A3})$$

By the chain rule,

$$\frac{d\tilde{\rho}}{dT} = \frac{d\tilde{\rho}}{dz} \frac{dz}{dT}. \quad (\text{A4})$$

The first term is

$$\frac{d\tilde{\rho}}{dz} = -\frac{3x}{2} \left( \frac{y - b^2}{z^2} \right) \left( \frac{y - b^2(1-z)}{z} \right)^{1/2} \quad (\text{A5})$$

and the second term is

$$\frac{dz}{dT} = \frac{\pi}{P} \sin \left[ \frac{2\pi T}{P} \right]. \quad (\text{A6})$$

Combining Equations (A2), (A4), (A5), and (A6) yields the exact Jacobian

$$\begin{aligned} J = \frac{d\tilde{\rho}}{dT} &= -\frac{6\pi^3}{P^3 G} \left( \frac{(1+r)^2 - b^2}{\sin^4[\pi T/P]} \right) \\ &\times \left( \frac{(1+r)^2 - b^2(1 - \sin^2[\pi T/P])}{\sin^2[\pi T/P]} \right)^{1/2} \\ &\times \sin \left[ \frac{2\pi T}{P} \right]. \end{aligned} \quad (\text{A7})$$

Making the small angle approximation  $\sin \phi \approx \phi$  (assuming  $\pi T \ll P$ ) and collecting terms yields

$$\begin{aligned} J &= -\frac{12\pi^3}{P^3 G} ((1+r)^2 - b^2) \\ &\times ((1+r)^2 - b^2(1 - [\pi T/P]^2))^{1/2} \left( \frac{\pi T}{P} \right)^{-4}. \end{aligned} \quad (\text{A8})$$

Once again taking advantage of  $\pi T \ll P$  simplifies the expression further to

$$J = -\frac{12\pi^3}{P^3 G} ((1+r)^2 - b^2)^{3/2} \left( \frac{\pi T}{P} \right)^{-4}. \quad (\text{A9})$$

### Appendix B Derivation of Jacobian for $T_{14} \rightarrow \ln \tilde{\rho}$

To derive the Jacobian of the transformation  $T \rightarrow \ln \tilde{\rho}$ , we note that

$$\frac{d \ln \tilde{\rho}}{dT} = \frac{1}{\tilde{\rho}} \frac{d\tilde{\rho}}{dT}. \quad (\text{B1})$$

Adopting our usual approximations  $\sin \phi \approx \phi$ ,  $\pi T \ll P$ , we may rewrite Equation (A1) in the simplified form

$$\tilde{\rho} \equiv \left( \frac{4\pi^2}{P^2 G} \right) ((1+r)^2 - b^2)^{3/2} \left( \frac{\pi T}{P} \right)^{-3}. \quad (\text{B2})$$

Combining Equations (A9), (B1), and (B2) and canceling terms yields

$$\frac{d \ln \tilde{\rho}}{dT} = -\frac{3}{T}. \quad (\text{B3})$$

We see that  $d \ln \tilde{\rho}/dT$  is independent of  $b$ .

### Appendix C Derivation of Jacobian for $T_{14} \rightarrow a/R$

To derive the Jacobian of the transformation  $T \rightarrow a/R_*$ , we define  $\alpha \equiv a/R_*$  and recognize that from Seager & Mallén-Ornelas (2003; their Equations (8) & (9)),

$$\alpha = \left( \frac{4\pi^2}{P^2 G} \right)^{-1/3} \tilde{\rho}^{1/3}. \quad (\text{C1})$$

By the chain rule,

$$\frac{d\alpha}{dT} = \frac{d\alpha}{d\tilde{\rho}} \frac{d\tilde{\rho}}{dT}. \quad (\text{C2})$$

The first term is

$$\frac{d\alpha}{d\tilde{\rho}} = \frac{1}{3} \left( \frac{4\pi^2}{P^2 G} \right)^{-1/3} \tilde{\rho}^{-2/3} \quad (\text{C3})$$

and the second term we derived previously. Adopting our usual approximations  $\sin \phi \approx \phi$ ,  $\pi T \ll P$  and combining Equations (A9), (B2), (C2), and (C3) yields

$$\frac{d\alpha}{dT} = -\frac{\pi}{P} ((1+r)^2 - b^2)^{1/2} \left( \frac{\pi T}{P} \right)^{-2}. \quad (\text{C4})$$

### Appendix D Derivation of Jacobian for $T_{14} \rightarrow \ln a/R_*$

To derive the Jacobian of the transformation  $T \rightarrow \ln a/R_*$ , we note that

$$\frac{d \ln \alpha}{dT} = \frac{1}{\alpha} \frac{d\alpha}{dT} \quad (\text{D1})$$





where as before  $\alpha \equiv a/R_*$ . Following our usual strategies and combining Equations (C1), (C4), and (D1), we arrive at

$$\frac{d \ln \alpha}{dT} = -\frac{1}{T}, \quad (\text{D2})$$

which is independent of  $b$ .

### ORCID iDs

Gregory J. Gilbert  <https://orcid.org/0000-0003-0742-1660>  
 Mason G. MacDougall  <https://orcid.org/0000-0003-2562-9043>  
 Erik A. Petigura  <https://orcid.org/0000-0003-0967-2893>

### References

- Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, *PASP*, **125**, 989
- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, **156**, 123
- Barnes, J. W. 2007, *PASP*, **119**, 986
- Berger, T. A., Huber, D., Gaidos, E., & van Saders, J. L. 2018, *ApJ*, **866**, 99
- Borucki, W. J., Koch, D., Basri, G., et al. 2010, *Sci*, **327**, 977
- Carter, J. A., Yee, J. C., Eastman, J., Gaudi, B. S., & Winn, J. N. 2008, *ApJ*, **689**, 499
- Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, *ApJL*, **529**, L45
- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2020, *AJ*, **160**, 159
- Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, *ApJS*, **224**, 12
- Crossfield, I. J. M., Petigura, E., Schlieder, J. E., et al. 2015, *ApJ*, **804**, 10
- David, T. J., Hillenbrand, L. A., Petigura, E. A., et al. 2016, *Natur*, **534**, 658
- Eastman, J. 2017, EXOFASTv2: Generalized publication-quality exoplanet modeling code, Astrophysics Source Code Library, ascl:1710.003
- Eastman, J., Gaudi, B. S., & Agol, E. 2013, *PASP*, **125**, 83
- Espinoza, N. 2018, *RNAAS*, **2**, 209
- Foreman-Mackey, D., Luger, R., Agol, E., et al. 2021, *JOSS*, **6**, 3285
- Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, *AJ*, **154**, 109
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, *A&A*, **616**, A1
- Gibson, S. R., Howard, A. W., Marcy, G. W., et al. 2016, *Proc. SPIE*, **9908**, 990870
- Gilbert, G. J. 2022, *AJ*, **163**, 111
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Natur*, **585**, 357
- Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2000, *ApJL*, **529**, L41
- Hoffman, M. D., & Gelman, A. 2011, arXiv:1111.4246
- Kipping, D. M. 2010a, *MNRAS*, **407**, 301
- Kipping, D. M. 2010b, *MNRAS*, **408**, 1758
- Kipping, D. M. 2014, *MNRAS*, **440**, 2164
- Kipping, D. M., & Sandford, E. 2016, *MNRAS*, **463**, 1323
- Kovács, G., Zucker, S., & Mazeh, T. 2002, *A&A*, **391**, 369
- Luger, R., Agol, E., Foreman-Mackey, D., et al. 2019, *AJ*, **157**, 64
- Mandel, K., & Agol, E. 2002, *ApJL*, **580**, L171
- Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, *ApJS*, **217**, 31
- NASA Exoplanet Archive 2021, Kepler Objects of Interest DR25, Version: 2021-08-07 13:20, NExSci-Caltech/IPAC, doi:10.26133/NEA5
- Neal, R. 2011, Handbook of Markov Chain Monte Carlo (Boca Raton, FL: CRC Press), 113
- Rauer, H., Catala, C., Aerts, C., et al. 2014, *ExA*, **38**, 249
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, *JATIS*, **1**, 014003
- Rowe, J. F., Bryson, S. T., Marcy, G. W., et al. 2014, *ApJ*, **784**, 45
- Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, *ApJS*, **217**, 16
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. 2016, *PeerJ Comput. Sci.*, **2**, e55
- Seager, S., & Mallén-Ornelas, G. 2003, *ApJ*, **585**, 1038
- Seifahrt, A., Stürmer, J., Bean, J. L., & Schwab, C. 2018, *Proc. SPIE*, **10702**, 107026D
- Stassun, K. G., Collins, K. A., & Gaudi, B. S. 2017, *AJ*, **153**, 136
- Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, *ApJS*, **235**, 38
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, **17**, 261
- Winn, J. N. 2010, in Exoplanets, ed. X. S. Seager (Tucson, AZ: University of Arizona Press), 55