UNIVERSITY OF CALIFORNIA
SANTA CRUZ

LEVERAGING THE POWER OF A PLANET POPULATION: MASS-RADIUS RELATION, HOST STAR MULTIPLICITY, AND COMPOSITION DISTRIBUTION OF *KEPLER*'S SUB-NEPTUNES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

ASTRONOMY & ASTROPHYSICS

by

Angie K. Wolfgang

June 2015

The Dissertation of Angie K. Wolfgang
is approved:

Professor Gregory P. Laughlin, Chair

Professor Jonathan J. Fortney

Doctor Natalie M. Batalha

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **vi**

List of Tables **vii**

Abstract **ix**

Dedication **xi**

Acknowledgments **xii**

1 Introduction **1**
1.1 The Discovery of Extrasolar Planets 1
1.2 Unearthing the Sub-Neptune Population: the *Kepler Mission* 4
1.2.1 From Photometry to Planets: Constructing the Catalog 6
1.2.2 Crucial Considerations: Survey Reliability and Completeness . . 9
1.2.3 Follow-Up Observations of *Kepler* Planet Candidates 13
1.3 From Discovery to Characterization of a Population 16

I Data-Driven Characterization of Sub-Neptune-Sized Planets **20**

2 Cross-Survey Comparison: *Kepler* and HARPS **21**
2.1 Introduction 21
2.2 The *Kepler* Data Set 26
2.3 The Transit-RV Comparison: Methods 27
2.3.1 Simulations of the Radial Velocity Population 28
2.3.2 Population-wide Mass-to-Radius Relationships 31
2.3.3 Star Selection 36
2.3.4 Detectability of Simulated Planets 39
2.3.5 Comparison with Total Number of True *Kepler* Planets 41
2.4 The Transit-RV Comparison: Results 43
2.4.1 Single-Valued M-R 43
2.4.2 Multi-Valued M-R 47

2.5 Discussion . . . 50
2.5.1 Caveats . . . 55
2.6 Conclusion . . . 60

3 Probabilistic Mass-Radius Relation for Sub-Neptune-Sized Planets 62
3.1 Introduction . . . 62
3.2 Modeling the M-R Relation . . . 65
3.3 Data . . . 68
3.4 Fitting the M-R Relations . . . 74
3.5 Results . . . 78
3.5.1 Deterministic vs. Probabilistic M-R Relations . . . 78
3.5.2 Changing the Dataset . . . 80
3.6 Discussion . . . 83
3.6.1 Visualizing the M-R Relation . . . 83
3.6.2 Using the M-R Relation to Predict Masses . . . 86
3.7 Conclusions . . . 87

4 Adaptive Optics Follow-up of *Kepler*'s Sub-Neptunes 89
4.1 Introduction . . . 89
4.1.1 Previous High-Resolution Imaging . . . 91
4.2 Target Selection . . . 94
4.3 Observations . . . 97
4.4 Data Reduction . . . 101
4.5 Results . . . 102
4.6 Further Work . . . 127

II Theory-Driven Characterization of *Kepler*'s Sub-Neptunes 130

5 The Composition Distribution of *Kepler*'s Sub-Neptunes 131
5.1 Introduction . . . 131
5.1.1 Modeling Sub-Neptune Interior Structures . . . 132
5.1.2 Statistical Treatment of Planet Populations . . . 136
5.2 Using the *Kepler* Objects of Interest . . . 137
5.2.1 Selecting a Complete Subsample . . . 138
5.3 Methods: Characterizing Planet Compositions via Statistical Modeling . 142
5.3.1 Choosing an Appropriate Statistical Framework . . . 144
5.3.2 Applying Bayes' Theorem . . . 148
5.3.3 Hierarchical Bayesian Modeling . . . 153
5.3.4 Our Hierarchical Model . . . 156
5.3.5 JAGS: MCMC with Hierarchical Models . . . 161
5.4 Results . . . 162
5.4.1 Population Composition Distribution . . . 162
5.4.2 Individual Planet Compositions . . . 165
5.4.3 Posterior Checks and Convergence . . . 177

5.5 Discussion . . . 182
5.5.1 Motivation for Salient Model Assumptions . . . 183
5.5.2 Radius as a Proxy for Composition . . . 186
5.5.3 The Rock-Gas Transition . . . 187
5.5.4 No Deterministic Mass-Radius Relationship . . . 191
5.5.5 Implications for Population Formation Models . . . 193
5.6 Conclusions . . . 194

6 Future Work **197**
6.1 Framework for the Analysis of Planet Populations . . . 199
6.2 Compositional Evolution of Sub-Neptune-Sized Planets . . . 200
6.2.1 Dependence of Composition on Incident Flux . . . 201
6.2.2 Population of Water-Dominated Planets . . . 203
6.3 Orbital Evolution of Sub-Neptune-Sized Planets . . . 205
6.3.1 Characterizing Distributions Relevant to Dynamics Studies . . . 206
6.3.2 Probabilistic analysis of Kozai-Lidov oscillations . . . 207

List of Figures

1.1 *Kepler*'s smallest planet candidates 5

2.1 *Kepler* Main Sequence Stellar Sample and Q0-Q2 Planet Detectability . 38
2.2 Single-valued M-R relation: example simulated period-radius distribution 43
2.3 Single-valued M-R relation: total number of detectable simulated planets 46
2.4 Multi-valued M-R relation: example simulated period-radius distribution 47
2.5 Multi-valued M-R relation: total number of detectable simulated planets 49

3.1 Graphical model of statistical relationships between individual observables and the mass-radius relation parameters 67
3.2 Posteriors for the parameters in our family of M-R relations 79
3.3 Posteriors for Eqn 3.2's M-R relation parameters for different datasets . 81
3.4 The dataset and the best-fit M-R relation. 84
3.5 M-R relation marginalized over the hyperparameter distributions 85

4.1 *Kepler* magnitude distribution of our AO sample vs. all planet candidates 98
4.2 Separation vs. magnitude difference of additional sources detected near *Kepler* Objects of Interest 103
4.3 Cumulative period distributions of the planets orbiting KOIs with or without visual companions 129

5.1 Radius distribution of all *Kepler* sub-Neptunes vs. our complete subsample 141
5.2 Signal-to-noise ratios of our complete subsample 142
5.3 Map of planet and star properties used to calculate planet compositions 146
5.4 Graphical model of statistical relationships between individual and population-wide parameters 158
5.5 Sub-Neptune composition distribution and its hyperparameters' posterior 163
5.6 Individual planet compositions as a function of radius 166
5.7 Assessing the hierarchical model: posterior vs. prior radius distributions 179
5.8 Convergence diagnostics for our hierarchical MCMC 182
5.9 The sample radius distribution colored by composition 185
5.10 The cumulative fraction of planets that are rocky 189
5.11 Masses and radii generated from our composition distribution 192

List of Tables

3.1 Masses and Radii of Small Planets 69
3.1 Masses and Radii of Small Planets 70
3.1 Masses and Radii of Small Planets 71
3.1 Masses and Radii of Small Planets 72
3.1 Masses and Radii of Small Planets 73
3.1 Masses and Radii of Small Planets 74
3.2 Best-Fit Parameters of the M-R Relation 78

4.1 Visual Companions Within 10″ 104
4.1 Visual Companions Within 10″ 105
4.1 Visual Companions Within 10″ 106
4.1 Visual Companions Within 10″ 107
4.1 Visual Companions Within 10″ 108
4.1 Visual Companions Within 10″ 109
4.1 Visual Companions Within 10″ 110
4.1 Visual Companions Within 10″ 111
4.1 Visual Companions Within 10″ 112
4.1 Visual Companions Within 10″ 113
4.1 Visual Companions Within 10″ 114
4.1 Visual Companions Within 10″ 115
4.1 Visual Companions Within 10″ 116
4.1 Visual Companions Within 10″ 117
4.1 Visual Companions Within 10″ 118
4.1 Visual Companions Within 10″ 119
4.1 Visual Companions Within 10″ 120
4.1 Visual Companions Within 10″ 121
4.1 Visual Companions Within 10″ 122
4.1 Visual Companions Within 10″ 123
4.1 Visual Companions Within 10″ 124
4.2 KOIs with No Visual Companions Within 10″ 124
4.2 KOIs with No Visual Companions Within 10″ 125
4.2 KOIs with No Visual Companions Within 10″ 126

4.2 KOIs with No Visual Companions Within 10″ 127

5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 167
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 168
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 169
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 170
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 171
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 172
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 173
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 174
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 175
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 176
5.1 Compositions of Individual Sub-Neptune-sized Planets in Sample . . . 177

Abstract

Leveraging the Power of a Planet Population:

Mass-Radius Relation, Host Star Multiplicity, and Composition Distribution of

Kepler's Sub-Neptunes

by

Angie K. Wolfgang

With the advent of large, dedicated planet hunting surveys, the search for extrasolar planets has evolved into an effort to understand the properties and formation of a planet population whose characteristics continue to surprise the provincial perspective we've derived from our own Solar System. The *Kepler Mission* in particular has enabled a large number of these studies, as it was designed to stare simultaneously at thousands of stars for several years and its automated transit search pipeline enables fairly uniform detection criteria and characterizable completeness and false positive rates. With the detection of nearly 5000 planet candidates, 80% of which are smaller than 4 $R_\oplus$, *Kepler* has especially illuminated the unexpectedly vast sub-Neptune population. Such a rich dataset provides an unprecedented opportunity for rigorous statistical study of the physics of these planets that have no analogs in our Solar System.

Contributing to this endeavor, I present the statistical characterization of several aspects of this population, including the comparison between *Kepler*'s planet candidates and low-mass occurrence rates inferred from radial velocity detections, the relationship between a sub-Neptune's mass and its radius, the frequency of *Kepler* planet

candidate host stars which have nearby visual companions as revealed by follow-up high resolution imaging, and the distribution of gaseous mass fractions that these sub-Neptunes could possess given a rock-plus-hydrogen composition. To do so, I have used sophisticated statistical analyses such as Monte Carlo simulations and hierarchical Bayesian modeling to tie theory more closely to observations and have acquired near infrared laser guide star adaptive optics imaging of 196 *Kepler* Objects of Interest. I find that even within this sub-Neptune population these planets are very diverse in nature: there is intrinsic scatter in masses at a given radius, the planet host stars have visual companions at a wide range of separations, and the composition distribution spans two orders of magnitude, with a peak at 1% hydrogen and helium by mass. There is much work to be done to explain this diversity quantitatively, and especially to tie these results to various planet formation scenarios; I have no doubt that many more surprises await us.

To my frequent existential crises,

for teaching me grit and self-forgiveness.

To the Power that connects us all,

for always guiding me back to my path and

towards a deeper understanding of my purpose.

Acknowledgments

First, I would like to thank my family, and especially my parents, for their unflagging belief in my potential and ability to not only finish but excel in whatever I set out to do. Their constant moral support in both this journey and the larger one of life means more to me than I could ever hope to express. I would not have made it this far without them, and I am forever, humbly grateful that I am their daughter.

I extend my sincerest thanks to the many people who encouraged me to pursue my Ph.D. and endowed me with the opportunities to do so, and to those who helped me survive grad school once I got there: to Prof. Itai Cohen, for enthusiastically welcoming me into his soft condensed matter lab at Cornell as a bright-eyed college freshman despite my lack of experience and coursework and for nominating me for many research programs and scholarships; to now Prof. Sharon Gerbode, for being the coolest, realest, most inquisitive, most brilliant, and seriously awesome physicist role model a girl could ever ask for; to now Prof. Jason Wright for being one of the most accessible, friendly, and genuinely caring advisors and mentors I've ever had, and for throwing the door wide open, taking me in, and really listening when I was deeply doubting myself and badly needed support and reassurance; to now Prof. Kevin Covey, for giving me an amazing REU experience at the CfA that convinced me that I really did want to go to grad school; to Prof. Jamie Lloyd for keeping me on as a junior researcher after that first summer when undergrad astronomy research positions were hard to come by, and for writing me grad school letters of recommendation; to Prof. Gregory Laughlin for serving as my graduate research advisor over the last six years, for writing me a

ton of recommendation letters, paying me over the summer, being open to funding any conference or travel I wanted to go on, and for allowing me to find my own research path; to Prof. Jonathan Fortney for his invaluable feedback on my performance, his advice about the interpersonal side of science, and for welcoming me into the Fortney group even when I wasn't officially his student; to Dr. Natalie Batalha for enthusiastically including me in *Kepler* business, serving on my committee, finding the time in her extremely busy schedule to write a seriously large number of postdoc recommendation letters, and for being an extremely successful female astronomer as I aspire to be; to my roommates, now Drs. Rosalie McGurk and Jerome Fang, for being hands-down the best people to live with for the last six years, and for being the kind of real friends who cry with you when you're sad and dance with you when you're happy; to Dr. Claire Dorman and soon-to-be Dr. Nathan Goldbaum, for our many nights of hanging out, commiserating, and laughing, and for making our cohort so awesome; to all of the UCSC astronomy graduate students who actively participate in our community, for providing much-needed distraction and fun and for helping me feel like I belonged; and to all who care about equity and inclusion in astronomy, for doing the very, very difficult, thankless, sometimes even unwanted work to make our profession more welcoming to people of all backgrounds, lived experiences, and orientations — the very people who we desperately need to add fresh perspective to our quest for scientific knowledge.

I would also like to thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for organizing a 3-week workshop on the Statistical Analysis of Kepler data in June 2013, during which I learned how to apply HBM and relevant computational

techniques to this thesis research, under the guidance of Merlise Clyde, Eric Ford, Jessi Cisewski, Robert Wolpert, and others. This opportunity was invaluable to me, as it finally provided a center to my thesis research and a clear path forward on which I was excited to embark. Additionally, I would like to thank the members of the Bayesian Characterization of Exoplanet Populations group which arose out of this workshop, for their invaluable advice and feedback, particularly Leslie Rogers, Tom Loredo, Darin Ragozzine, Megan Shabram, Daniel Foreman-Mackey, and David Hogg, in addition to those listed above. The sense of community in this group is strong and gave me the support, confidence, and motivation to keep going when I felt discouraged and my outlook on my work was bleak.

My financial support during my graduate work was generously awarded to me by the UCSC Graduate Division's Eugene Cota-Robles Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. 0809125. These fellowships allowed me the freedom to pursue my own avenues of scientific inquiry, and I strongly believe that I am a better scientist for it. Additional summer funding was provided by NASA contract NNG14FC03C through subaward agreement 5710003702 and by the NASA Ames NAI Astrobiology Module (GL).

The material in Chapters 3 and 5 was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute, with some minor support from the NSF Grant No. PHY11-25915 to the UCSB Kavli Institute for Theoretical Physics, which brought together the authors of Chapter 3 during its Dynamics and Evolution of Earth-like Plan-

ets program. 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.

This research is based on data collected by the Kepler mission. Funding for the Kepler mission is provided by the NASA Science Mission directorate. This research has also made extensive 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. Additional data were obtained from the Multimission Archive at the Space Telescope Science Institute (MAST); PyRAF is also a product of the Space Telescope Science Institute. 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 NNX09AF08G and by other grants and contracts. Finally, this research has made use of the Exoplanet Orbit Database and the Exoplanet Data Explorer at exoplanets.org.

Chapters 2, 3, and 5 of this dissertation includes reprints of the following previously published material, respectively: Wolfgang, A. & Laughlin G. "The Effect of Population-wide Mass-to-radius Relationships on the Interpretation of *Kepler* and HARPS Super-Earth Occurrence Rates", Astrophysical Journal, 750, 148 (2012); Wolfgang, A., Rogers, L. A., & Ford, E. B. "Probabilistic Mass-Radius Relationship for Sub-Neptune-Sized Planets", Astrophysical Journal, submitted, arXiv:1504.07557 (2015); and Wolfgang, A. & Lopez, E. "How Rocky Are They? The Composition Dis-

tribution of *Kepler*'s Sub-Neptune Planet Candidates within 0.15 AU", Astrophysical Journal, in press, arXiv:1409.2982 (2015). The coauthor listed in the first publication directed and supervised the research which formed the initial basis for the dissertation. The coauthors for the last two publications grant their permission for the material to be used in this dissertation; in both cases the primary author assembled and selected the data, developed the final version of the presented statistical model (with sole development from beginning to end for the last article), ran the described simulations, wrote virtually all of the text, made the figures, and served as primary contact for the referee and publication process.

Chapter 1

Introduction

1.1 The Discovery of Extrasolar Planets

When I was born 28 years ago, humanity knew of 9 planets: Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and now dwarf planet Pluto. Today we know of thousands (Akeson et al., 2013; Han et al., 2014). Thanks to these discoveries, the last two decades have marked a revolution in our understanding of the "typical" planet produced by the formation and evolution processes in protoplanetary disks, and the implications are, simply put, astronomical. First, we do not really understand how planets form: each new technology and detection technique that we apply to our search reveals unexpected planets with properties that surprise us. Second, the sheer number of extrasolar planets is astounding, fundamentally shifting our perspective on our uniqueness and place in the Universe. And third, we have a lot yet to learn about both these planets and the ones that have yet to be discovered. The field of exoplanetary science is young, but its future is bright.

While the first observational evidence for the existence of extrasolar planets rests with the super-Jupiters discovered by Campbell et al. (1988) and Latham et al. (1989), our story of coming to expect the unexpected really began with PSR B1257+12 b, c, and d (Wolszczan & Frail, 1992). When precise timing of the pulsar revealed periodic variations indicative of planetary mass bodies, we received our first hint of the diversity of extrasolar planets. Soon afterward, 51 Peg b (Mayor & Queloz, 1995) heralded the population of "hot Jupiters" whose tight orbits challenged planet formation theories that had been developed to explain the architecture of our own Solar System. Concerted efforts over the subsequent decade to push to smaller and lower-mass planets yielded first GJ 876 d (Rivera et al., 2005) with the radial velocity detection method and then eventually the transiting CoRoT-7b (Queloz et al., 2009; Léger et al., 2009) and GJ 1214b (Charbonneau et al., 2009), the first examples of the super-Earths and sub-Neptunes class of planets that is the subject of this thesis. Indeed, these discoveries foreshadowed the momentous success of the *Kepler Mission* in unearthing yet another unexpected planet population: the thousands of sub-Neptune-sized planets to which there are no Solar System analogues (Mullally et al., 2015).

To understand the opportunities and limitations for further investigations of these exoplanets, it is necessary to first overview the detection techniques that found them. The large majority of the first decade of exoplanet discoveries were made by the radial velocity (RV) technique: 120 of the 128 extrasolar planets discovered before 2005 were first detected this way (Han et al., 2014). Conceptually, this is very similar to the well-established method of searching for close binary stars by detecting sinusoidal

(or otherwise Keplerian) Doppler shifts that are due to the target star's radial reflex motion induced by its companion's gravitational force. To push to lower masses, the planet-hunting RV method uses echelle spectroscopy over a very large wavelength range and the simultaneous observation of a wavelength calibrator to obtain < 3 m/s^{-1} RV precision, below the $\sim 10 - 50$ m s^{-1} floor that had previously limited this technique to detecting more massive companions (Butler et al., 1996); the use of a fiber on the front end of the spectrograph also provides necessary stability to achieve high RV precisions (Queloz et al., 1999; Pepe et al., 2000). After a sufficient observing time baseline has been achieved to adequately cover enough of the orbital phase space, then the period, minimum mass $m\sin(i)$, and eccentricity of the planetary companion can be measured.

The transit detection technique, on the other hand, took several more years to yield detections, beginning with HD 209458 b (Charbonneau et al., 2000). This discovery justified existing efforts and instigated new work to develop serious transit search surveys (i.e. Udalski et al., 2002; Bakos et al., 2002; Alonso et al., 2004; Pollacco et al., 2006; Nutzman & Charbonneau, 2008; Borucki et al., 2010), which seek to find periodic drops in the target star's flux that indicate some object is blocking out a portion of the host star's light. Again, this is conceptually similar to the detection of eclipsing binary stars, but improvements in photometric precision and observing duty cycle and automation were needed to find the small ($< 1\%$ drop in flux), hour-long signals among days' worth of observations. When multiple transits are detected, they enable measurement of the planetary radius (provided the stellar radius is well known), the period of the orbit, and the resolution of the $\sin(i)$ degeneracy in any RV mass

measurements that can be obtained, as the transit geometry places tight constraints on the orbital inclination.

1.2 Unearthing the Sub-Neptune Population: the *Kepler Mission*

The transit search detection technique had found a few dozen planets by the time the *Kepler Mission* came online in mid-2009 (Han et al., 2014), setting the stage for the truly astounding discoveries that *Kepler* would make. A 4-year search for potentially habitable Earth-sized planets around solar-type stars (Borucki et al., 2010), *Kepler* continuously monitored $\sim$ 200,000 stars on a 30-minute cadence in its > 100 deg^2 field of view (Koch et al., 2010) for periodic photometric dips that fit the shape and duration of a planetary transit. The telescope's $\sim 0.01''$ per hour pointing stability (Koch et al., 2010) and 10-100 ppm photometric precision (Jenkins et al., 2010a) enabled, for the first time in the short history of exoplanet searches, the detection of planets that are Earth-sized or smaller. Furthermore, its Earth-trailing heliocentric orbit facilitated continuous data acquisition on a single patch of sky without the diurnal or annual cycles that generate aliases (Koch et al., 2010), enabling the detection of long-period planets out to year-long orbits.

Due in part to these design specifications, *Kepler* has produced a veritable cornucopia of planet detections. Over the 4-year lifetime of the mission, it discovered 4604 planet candidates (Exoplanet Archive [Akeson et al. 2013], cumulative table including planet candidates from the latest search [DR24] through all quarters of data [Q1-17]).

3730 of these are smaller than the size of Neptune ($R_{Nep} = 4$ $R_\oplus$), as displayed in Figure 1.1, and 865 of those are "confirmed" with very high statistical likelihood of being bona-fide planets (see §1.2.2 for a discussion of this). With a productivity unrivaled by nearly every other planet search effort, *Kepler* has thrown the door wide open on a population of planets that two decades ago we had no way of knowing existed: the super-Earths and sub-Neptunes — those exoplanets with 1 $R_\oplus < R_{pl} < 4$ $R_\oplus$— which have no analogs in our Solar System.

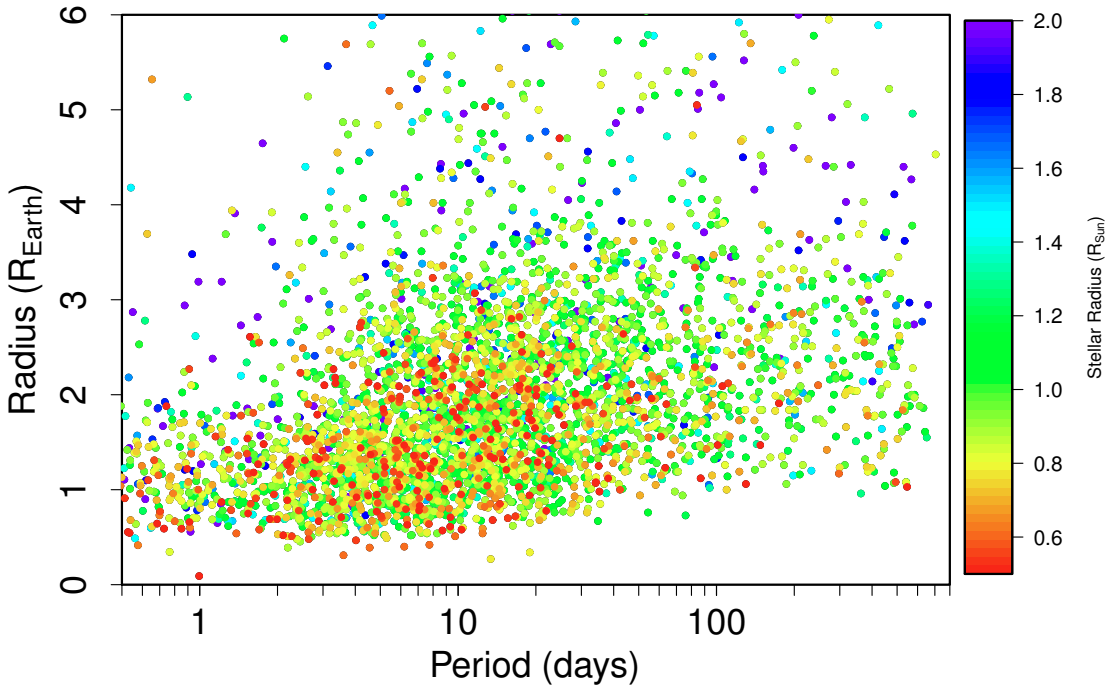

Figure 1.1: Periods and radii of *Kepler* planet candidates with $R_{pl} < 6$ $R_\oplus$, from the latest data release (Q1-17, DR24); those with $R_{pl} < 4$ $R_\oplus$ constitute the "sub-Neptune" population that provides the observational foundation of this thesis. The points are color coded according to the size of the host star, as reported on the NASA Exoplanet Archive; stars which fall below or above the color scale are colored red or purple, respectively.

This unprecedented sample size has enabled a plethora of statistical studies of these planets' properties, as overviewed in §1.3. However, the development of this planet catalog is only possible with a significant amount of automated processing, extensive human inspection of the resulting transit signals, and community-driven follow-up efforts encompassing many hours of ground-based telescope time. A detailed understanding, and subsequent treatment, of the complexity of the *Kepler* data is therefore absolutely necessary before any statistical study that uses this catalog can be sure its conclusions reflect the properties of the underlying population and not the detection biases or selection effects which have shaped the catalog as currently reported. We outline this processing and the discuss the need for observational follow-up in the following subsections.

1.2.1 From Photometry to Planets: Constructing the Catalog

To transform the thousands of pixel photon counts transmitted to Earth each month by the *Kepler* spacecraft into a list of possible exoplanets is an enormous undertaking. Much of the initial processing is done by an extensive automated pipeline developed over many years by the *Kepler Mission* Science Operations Center (SOC) and which experienced continued evolution throughout the *Kepler* mission to optimize the detection of true planetary transits. This software is outlined in Jenkins et al. (2010b); it includes modules that calibrate the raw images from the *Kepler* spacecraft (Quintana et al., 2010), produce raw photometry from these images (Twicken et al., 2010b), detrend the raw photometric time series to remove systematic instrumental effects (Twicken et al., 2010a), search for transit-like signals using a wavelet-based

adaptive-matched filter (Jenkins et al., 2010c), and then fit a transit model to further characterize the signal (Wu et al., 2010).

The human effort continues where this automated pipeline ends: with the list of Threshold Crossing Events (TCEs), the periodic transit-like signals which pass the pipeline's filter. The number of TCEs can be substantial, on the order of 10^4 (Tenenbaum et al., 2013, 2014; Seader et al., 2015), and a large number are produced by astrophysical variability or instrumental noise ($> 80\%$ of Q1-8 TCEs; Burke et al. 2014). To identify these cases, the TCE list undergoes a "triage" stage where the Data Validation (DV) summary pages produced by the last module in the SOC pipeline are glanced over by the Threshold Crossing Event Review Team (TCERT)[1], of which I am a member. If the shape of the signal has the characteristic U- or V-shape of a planetary transit or stellar eclipse, it is made into a *Kepler* Object of Interest (KOI).

The next stage of human inspection involves determining whether a KOI is either a false positive (FP) or a planetary candidate (PC), a process referred to as "dispositioning". This step is necessary because there are a number of astrophysical scenarios that can produce photometric signals that are qualitatively similar to planetary transits. It can be difficult for automated detection algorithms to reliably and accurately distinguish between large mass ratio eclipsing binaries and planetary transits as well as other less common scenarios, yet they can be fairly quickly categorized by a human considering various aspects of the data simultaneously with complex, physics-based logic that was difficult to anticipate and preemptively automate at the start of the mission.

[1]The DV summary for every TCE produced since the Q1-12 pipeline run is available at the Exoplanet Archive.

The data and metrics used in this dispositioning process are explained in detail in Batalha et al. (2010a), Batalha et al. (2013), Burke et al. (2014), Rowe et al. (2015), and Mullally et al. (2015). In general, dispositioning involves at least two TCERT members reviewing the DV summaries in detail. These documents contain several plots and quantitative metrics that are helpful for identifying deep secondary eclipses indicative of stellar binaries, other out-of-transit photometric signatures, per-quarter variation in the photometry suggesting contamination from other sources, and statistically significant centroid offsets which place the location of the transit signal on another star. More recent versions of the vetting reports[2] used for TCERT dispositioning include detailed pixel-level difference images to assess the quality of and systematics in the centroid offset measurement, measures of the red noise in the light curve to quantify the false alarm rate[3], and light curves from alternate detrending algorithms to test the robustness of the transit signal. If the two TCERT members disagree in their dispositions, then the decision is made by a third or fourth "master" TCERT member. The resulting list of PCs is compared to itself and the list of known eclipsing binaries to identify recurring periods and epochs which indicate signal contamination on the CCD (see Coughlin et al. 2014 for details); if a match is identified, it is labeled as a FP. The KOIs and their dispositions are then uploaded to the Exoplanet Archive as one of the various KOI catalogs.

While the hope is that incorporating human decisions in this process increases

[2]These reports are also available on the Exoplanet Archive.

[3]As opposed to the false positive rate, which is the proportion of transit-like signals produced by any astrophysical eclipsing scenario other than a planet transiting the specified target star, the false alarm rate is the proportion of transit-like signals that are produced by the coincidental alignment of several-hour-long negative excursions in the data due to red noise; this is particularly a problem for the lowest SNR transit detections.

the quality of the planet candidate catalog, in practice there is some variability in decision-making from person to person (see Mullally et al. 2015 for an analysis of this effect). To improve the reproducibility of this process and reduce the significant man-hours involved, the *Kepler* team has begun to transition to automated vetting of the TCEs using several complex decision trees which are heavily based on the experience of TCERT. This automated vetting was implemented for the Q1-17, DR24 KOI catalog, the latest KOI catalog as of the time of this writing, with spot-checks provided by the TCERT team to tweak parameter values which increase the reliability and completeness of the resulting catalog.

1.2.2 Crucial Considerations: Survey Reliability and Completeness

Two key characteristics of the planet candidate/false positive dispositioning process overviewed in §1.2.1 are that it only considers data obtained by the *Kepler* spacecraft and that KOIs are considered "innocent until proven guilty". This inclusive approach therefore includes many of the more ambiguous planet candidates which can be statistically determined as false positives with additional modeling[4], or proven to not be planetary with additional follow-up data. False positive scenarios that are especially difficult to rule out from the *Kepler* data alone include background eclipsing binaries where chance alignments between a target star and a distant EB produce eclipse depths that are diluted by the foreground target to $\sim 0.01\%$, hierarchical triple systems with one or more eclipsing components (Gautier et al., 2010; Morton & Johnson, 2011b;

[4]In particular, a large transit depth or a V-shaped transit by itself is not considered sufficient evidence for a false positive disposition.

Morton, 2012), and even eccentric stellar binaries oriented such that only the secondary eclipse occurs (Santerne et al., 2013).

A number of studies have analyzed the effect of this inclusive, *Kepler* data-only approach on the reliability of the PC catalog. Morton & Johnson (2011b) found that overall the false positive probability (FPP) of KOIs labeled as planetary candidates are low ($\sim$ 5%); however, this FPP does increase to $\sim$ 10% for PCs with larger transit depths and increases with decreasing galactic latitude and apparent magnitude of the host star. Given the dependence of stellar multiplicity on the mass of the primary, one would expect that the FPP also increases as the stellar mass increases; indeed for giant star hosts, the FPP is above 50% (Sliski & Kipping, 2014). The analysis of Fressin et al. (2013) supports the finding of a low FPP for the smallest planets with a more detailed treatment of the unknown planet occurrence rate, finding FPP above 15% for PCs with radii $> 4R_{\oplus}$ as opposed to FPP $\sim$ 8% for the smaller Super-Earths.

Given these studies, the sub-Neptune-sized planets around dwarf stars that we consider here are the least likely of all the *Kepler* planet candidates to be false positives. This is especially true if those planets exist in multiple planet systems, as the probability that any one target star would be associated with multiple astrophysical false positive signals is particularly low (Lissauer et al., 2012, 2014). Rowe et al. (2014) used this fact to statistically validate over 700 *Kepler* PCs as true planets, most of which were sub-Neptune-sized due to the observed preference for these small planets to occur in multiple planet systems. Therefore, we proceed with our analyses without a FPP correction, as the statistical samples of *Kepler* planets that we use here are reliable at the 90-95%

level and therefore, in a population-wide sense, reliability is not a high-priority concern for sub-Neptunes.

A more significant problem is the the converse issue of completeness: how many planets exist in nature which *Kepler* did not detect, and how this sensitivity varies as a function of planet radius, period, stellar radius, etc. This detection bias has a direct effect on the observed planet candidate catalog: it causes the distributions of observed planet properties to substantially differ from those in the underlying population. If not corrected for, this survey incompleteness can therefore suggest incorrect conclusions about the planet population. As explained below, detection bias is especially problematic for the smallest planets in the catalog, the Earth- and super-Earth-sized planets, and so we take special care to account for this effect in this research.

Kepler's incompleteness arises from a number of different effects. The most obvious is arguably the transit probability: because a planet is more likely to transit its host star when the radius of the star is larger and the line-of-sight planet-star distance is smaller, *Kepler* will naturally detect more short period planets, planets around smaller stars, and, to second order, higher eccentricity planets (Burke, 2008; Kipping, 2014) . This effect must be corrected if the planet property of interest is expected to correlate with stellar type, planet period, or planet eccentricity.

In addition, *Kepler* is not able to detect every transiting planet in its field of view, as the stellar noise profile varies strongly from star to star. Given the range of noise levels across the target star sample, this leads to detecting a lower fraction of existing transiting planets at smaller radii. It is also more difficult to detect longer

period planets, as the more transits that occur, the more data that can be binned to beat down the noise. Fortunately, the Combined Differential Photometric Precision (CDPP) calculated by the *Kepler* pipeline (Christiansen et al., 2012), provides an estimate of the stellar noise seen by the pipeline and can be used to estimate the detectability of individual planets. Either implementing a forward model which includes this per-star, per-planet detection threshold as was done in Chapter 2 (published as Wolfgang & Laughlin 2012) or limiting the considered target star sample to the least noisy stars as was done in Chapter 5 (published as Wolfgang & Lopez 2015) can diminish both of these effects.

Finally, Batalha et al. (2013) showed that the *Kepler* pipeline outlined above detects fewer low signal-to-noise (SNR) transit-like signals than expected even when the stellar noise detection threshold is accounted for, pointing to yet another detection bias against small, long-period planets for which the SNR is particularly low. This pipeline incompleteness has only been recently realized to have a significant effect, and work to characterize it via transit injection is ongoing. First results presented in Christiansen et al. (2013) indicate that the later modules of the pipeline do not systematically perturb the signal strength of individual events in an individual quarter of data, although larger perturbations are measured for lower signal-to-noise events. On the other hand, studies over multiple quarters shows that the transit search module culls the lower signal-to-noise events to a significant degree, due in large part to metrics that have been implemented to discard false alarm detections (for some discussion of the inclusion and effect of the "veto" metrics used in the pipeline, see both Tenenbaum et al. 2014

and Seader et al. 2015). This aggressive rejection results in significant incompleteness ($< 95\%$) for transits with a phased and folded SNR < 15 (Seader et al., 2015). Without detailed transit injection results at the time of the work of this thesis, we could not fully correct for this effect, but instead offer qualitative comments about how results may change when it is quantitatively accounted for.

1.2.3 Follow-Up Observations of *Kepler* Planet Candidates

Following up *Kepler* Objects of Interest (KOIs) with a variety of ground-based observations is useful in several respects. First, the properties of the *Kepler* target stars were characterized via photometry at the start of the mission (Brown et al., 2011), and so spectroscopic follow-up can help constrain host stellar properties to greater precision[5]. This is especially important for studies which involve the *Kepler* radius distribution, such as that in Chapter 5: because *Kepler* actually measures the depth of the transit, and this depth is the ratio of the planetary to stellar surface areas, the precision of the planet's radius measurement is directly proportional to our knowledge of the host star's radius. As Huber et al. (2014) find in their collation of the most up-to-date *Kepler* target star observations and their subsequent uniform modeling of the stellar properties, spectroscopy can improve stellar radius constraints by up to a factor of 10.

Second, obtaining more information about these KOIs using other techniques and instruments is a straightforward approach to false positive identification that complements the *Kepler* data-only studies detailed in §1.2.2. There are numerous facets to

[5]The most precise stellar properties are offered by asteroseismology, which uses the *Kepler* data itself; however, only a few dozen of the thousands of KOIs exhibit the stellar oscillations needed for this technique to work (Huber et al., 2013).

this effort to identify bona-fide planets among the *Kepler* candidates, many of which are listed in Gautier et al. (2010). The spectroscopy and imaging involved in this effort, along with multi-band in-transit photometry that can identify stellar blends via wavelength-dependent transit depths (Désert et al., 2015), are useful even when they don't immediately rule out the KOI as a false positive. In the case of a non-detection, which supports the KOI's planetary nature, the observations can be incorporated into statistical planet validation software such as BLENDER (Torres et al., 2011), PASTIS (Díaz et al., 2014; Santerne et al., 2015), or VALFAST (Morton et al., 2014). These codes calculate the likelihood of the observations as well as the shape of the transit signal given various astrophysical false positive scenarios, and outputs the posteriori probability that the signal is due to a planetary transit. If this probability is above some threshold, the KOI is "validated" and considered a true planet.

The *Kepler* follow-up program can be summarized as the acquisition of two different kinds of observations: medium to high resolution spectroscopy, and high resolution imaging. The former is usually used to distinguish some of the astrophysical false positive scenarios that involve configurations of multiple stars, either through the measurement of large stellar radial velocity offsets over time, or for a smaller subset of these configurations, the identification of double-peaked lines or shifted spectral line bisectors (Gautier et al., 2010). Indeed, such radial velocity follow-up has proved invaluable for identifying a large number of false positives among the largest planets: Santerne et al. (2012) finds that 35% of the deepest, shortest period PC signals have masses too high for planetary bodies. Unfortunately, however, *Kepler*'s target stars consist pri-

marily of 14th and 15th magnitude G dwarfs (Batalha et al., 2010b), and so obtaining high-precision RV measurements proves prohibitively expensive for a majority of these planet candidates. As described in Chapters 2, 3, and 5, the faintness of the average *Kepler* target also proves problematic for detailed characterization of the planets' compositions. While this reality of the *Kepler* sample motivated the bright-star emphasis of K2 and the future TESS mission, it also motivates every chapter of this thesis research: both the adaptive optics imaging of *Kepler* host stars, and the implementation of more sophisticated statistical techniques to characterize the natures of these sub-Neptunes.

The need for high resolution imaging follow-up of *Kepler* targets arises from the large spatial coverage of an individual pixel on the photometer's CCD. Due to the engineering trade-offs between the spacecraft-to-ground data transfer bandwidth and the number of target stars that needed to be continuously observed to reach projected science goals, *Kepler*'s pixels were designed to be 4 across (Koch et al., 2010). While well motivated, this decision makes precise centroid positions difficult to measure; coupled with errors in the *Kepler* pixel response function (Bryson et al., 2010), this results in a median 3σ upper limit of $\sim 0.5''$ for the detection of statistically significant quarter-averaged centroids that are offset from the target star, with significant variability from target to target (standard deviation: $0.9''$; Bryson et al. 2013). Even excluding the presence of additional sources outside a radius of $2''$, a resolution that is effortlessly achieved with adaptive optics imaging on even modestly sized telescopes, is extremely helpful (Morton & Johnson, 2011b; Morton, 2012). Finally, measuring the relative brightnesses of nearby sources that could have contaminated the *Kepler*

lightcurve enables a correction to the diluted observed transit depth, thereby providing more accurate planetary parameters. Because AO imaging a particularly efficient use of observational resources given that only a few minutes on a ground-based telescope is required per target (versus several to many nights of high resolution spectroscopy to calculate radial velocities), obtaining these data offer a high-impact opportunity for the exoplanet community to contribute to this effort, as I have done in Chapter 4.

1.3 From Discovery to Characterization of a Population

The past decade has seen an extraordinary increase in our discovery of short-period extrasolar planets, due to the transit and radial velocity (RV) detection methods overviewed in §1.1, and especially to the *Kepler Mission* described in §1.2. Now with over 1500 confirmed planets and 3000 additional *Kepler* planet candidates (Akeson et al., 2013; Han et al., 2014), the state of the art has progressed to the point where statistical studies of entire planet populations are realistically feasible. The field of exoplanetary science is truly transitioning from an era of discovery to an era of characterization — the effort to study the properties and explain the origins of the planet populations that we have found.

The discovery of *Kepler*'s unexpected sub-Neptune-sized population in particular poses many questions about the fundamental properties of planets, such as the range of compositions that planet formation produces, how often they occur in multiple-planet systems, and what physics drives differences in their orbital architectures. The answers have profound implications for how these planets formed and evolved from their birth

conditions, and enable us to place our Solar System into an appropriate cosmic context. For example, these planets' compositions have been the subject of much scrutiny (e.g. Rogers et al., 2011; Valencia et al., 2013; Lopez & Fortney, 2014; Howe et al., 2014), as their sizes span the gap between the rocky and non-rocky planets of our Solar System. However, the vast majority of these studies have been focused on constraining such physical characteristics for individual planets. Given that over 3500 sub-Neptune-sized planet candidates have been discovered to date (Mullally et al., 2015), we must maximally leverage the statistical power of the entire population to answer compelling questions about these planets and their origins. A crucial part of this endeavor is therefore the recognition of the wide range of statistical techniques that have been developed for such population studies and the careful choice of those that are most appropriate to our purposes.

Historically, radial velocity (RV) surveys provided the first opportunity to study the characteristics of an emerging population of planets. Tabachnik & Tremaine (2002) and Cumming et al. (2008) used a maximum likelihood approach with Poisson statistics to infer the joint mass-period distribution of detected RV planets, while necessarily incorporating RV detection thresholds to account for incompleteness (known as "truncation" in the statistical literature). More recently, Howard et al. (2010), Wittenmyer et al. (2011), and Mayor et al. (2011) extended the analysis of survey incompleteness to smaller masses, using a simple efficiency correction to obtain binned estimates of the occurrence rate of RV planets. The eccentricity distribution of RV planets were studied by Jurić & Tremaine (2008) and Ford & Rasio (2008), among others, who performed

K-S tests to compare the eccentricity distributions of their planet-planet scattering simulations to various sub-populations of RV planets to illuminate their origins. Hogg et al. (2010) detailed a hierarchical Bayesian analysis of the RV eccentricity distribution, providing the first exoplanet-specific example of the framework that we use in Chapters 3 and 5 of this thesis.

These population studies expanded in scope and feasibility with the advent of the *Kepler Mission*. Among them are analyses of the occurrence rate of exoplanets at different sizes and periods (Catanzarite & Shao, 2011; Youdin, 2011; Howard et al., 2012; Dong & Zhu, 2013; Fressin et al., 2013; Dressing & Charbonneau, 2013; Petigura et al., 2013; Morton & Swift, 2014; Foreman-Mackey et al., 2014), the consistency between RV and transit surveys (Wolfgang & Laughlin, 2012; Wright et al., 2012; Figueira et al., 2012), the *Kepler* eccentricity distribution (Moorhead et al., 2011; Kane et al., 2012; Wu & Lithwick, 2013), and the frequency and architecture of multiple-planet systems (Latham et al., 2011; Lissauer et al., 2011b; Tremaine & Dong, 2012; Fabrycky et al., 2014). The relationship between the radii and masses for sub-Neptune planets has also received attention (Wolfgang & Laughlin, 2012; Wu & Lithwick, 2013; Weiss & Marcy, 2014), with the aim of illuminating the compositions of these planets Rogers (2015); Wolfgang & Lopez (2015).

This thesis marks a significant contribution to this sub-Neptune population characterization effort, consisting of two refereed publications (Wolfgang & Laughlin 2012; Wolfgang & Lopez 2015; Chapters 2 and 5, respectively), another article that has been submitted (Chapter 3), and a sample of 200 near-infrared adaptive optics images of

Kepler planet hosts, one of the largest such samples at those wavelengths (Chapter 4). My focus has mainly been on topics relevant to the compositions of these sub-Neptune-sized planets: Chapter 2 describes a forward model that I constructed to calculate the number of planets drawn from the RV sub-Neptune population that would have been detectable by *Kepler* , a crucial step of which is consideration of the relationship between a sub-Neptune's mass and its radius; Chapter 3 constrains this mass-radius relationship directly with newly available transiting planet mass measurements and with the sophisticated statistical framework of hierarchical Bayesian modeling (HBM); and Chapter 5 implements theoretical internal structure models directly into this HBM framework to derive a composition distribution for these sub-Neptunes. However, the potential for characterization of planet populations extends far beyond that of compositions, and I will be pursing topics relevant to the origins of these sub-Neptunes for my postdoctorate work, as I detail in Chapter 6. This thesis establishes the foundation for this research, but there is still much to be done, especially with respect to tying planet population formation and evolution theory directly to the sub-Neptune detections and observations. Fortunately, the field of statistics provides us with the means to bridge this gap; it is with this eye on the most appropriate statistical tools to enable the theory-observation comparison that I endeavor to better understand these planets and how they came to be.

Part I

Data-Driven Characterization of Sub-Neptune-Sized Planets

Chapter 2

Cross-Survey Comparison: *Kepler* and HARPS

2.1 Introduction

As outlined in §1.3, the numerous planet discoveries made by transit and radial velocity surveys have enabled a large number of statistical surveys which provide insight into extrasolar planetary origins. A substantial challenge, however, still lies in synthesizing the results from different surveys into a cohesive picture of the Galactic planetary population, as each technique provides different information about the planets' physical characteristics and is subject to different selection biases.

These cross-survey considerations are especially important when one tries to compare the results of Doppler velocity surveys with the results of photometric transit surveys. Not only do these two detection methods generally sample different regions

of the Galaxy, but they also implement different observing strategies due to the intrinsically low geometric probability of a planetary transit and to the strict spectroscopic requirements needed to achieve 1 m/s precision in RV (see for example Rupprecht et al., 2004; Borucki et al., 2010; Koch et al., 2010; Batalha et al., 2010b). The result of these fundamental differences is that most RV-detected planets don't transit, and that most transiting planets suffer from a dearth of high-precision Doppler follow-up measurements. All is not lost, however: if these biases are properly accounted for, then one can utilize the global properties of the two samples to draw conclusions about the Galactic distribution of planetary properties.

A particularly valuable outcome of the transit-RV comparison originates from the distinction between measuring a planet's radius via a transit and measuring its mass via RV observations. When a Doppler-characterized planet is also observed to transit, these two quantities enable the range of possible compositions to be modeled even in the absence of any other observational constraints, through individual mass-to-radius relationships (M-Rs) calculated for a variety of interior planetary structures (e.g. Fortney et al., 2007; Valencia et al., 2007a; Seager et al., 2007; Rogers et al., 2011). However, because planets that are well-characterized by both methods are rare, one must apply *population-wide* M-Rs to make some headway in understanding the compositional distribution of planet populations. The use of these broad-brush M-Rs necessitates the assumption that transit and RV surveys adequately sample the full range and frequency of naturally occurring planetary compositions, once the populations are corrected for selection bias.

Before these population-wide M-Rs can be properly interpreted, it is essential to understand how they are fundamentally different from the M-Rs that are calculated using structural models of individual planets. On the most basic level, the transformation of a population of planetary masses to radii requires that a range of compositions be included *a priori*. While it is certainly true that inferring an individual planet's composition from its mass and radius is a degenerate problem and results in a range of possible part-iron, part-silicate, part-gas compositions, the bulk density of a planet, $\rho(M, R)$, is nonetheless typically well known from observations. This information is absent, however, when one compares a transiting planet population to a Doppler-detected population and the two samples have very few planets in common. As a result, bulk density is essentially a free parameter in transit-RV comparisons, and some assumptions about it, or about the compositions which correspond to it, must be made.

A key issue for the transit-RV comparison is how one chooses to parameterize planetary composition over the entire population. The simplest case would be if all planets had the same composition, as this enables the planets' masses to be straightforwardly converted to radii. However, Howard et al. (2012) have already shown that this very simple M-R fails to match the *Kepler* planet candidates when a power law is used for the planetary mass distribution, and so we consider more flexible and more physically motivated M-Rs in this chapter.

The plethora of ongoing planet searches enables the Galactic planetary census to be illuminated in a number of different ways. Two of the most influential surveys to date are the *Kepler Mission*, which found 1,235 transiting planet candidates in its

first four months of data (Borucki et al., 2011), and the Geneva High Accuracy Radial velocity Planet Search (HARPS), which has discovered over 85 planets during the course of observing hundreds of the brightest stars in the solar neighborhood (Ségransan et al., 2011). Both of these surveys are in a position to unearth the population of low-mass short-period planets and to provide statistics about their relative frequency. The initial results hint suggestively at the prevalence of truly Earth-like planets and are of particular interest for planet formation theories that strive to explain or predict the mass-distance distribution of planet populations (Ida & Lin, 2004; Kornet & Wolf, 2006; Schlaufman et al., 2009; Mordasini et al., 2009; Ida & Lin, 2010; Alibert et al., 2011).

Alarmingly, the low-mass planet occurrence rates measured by the two surveys appear to conflict with one another. Systematic statistical analyses of the short-period *Kepler* planet candidates have yielded 0.130 ± 0.008 planets per solar-type star (Howard et al., 2012) or 0.19 planets per solar-type star (Youdin, 2011), with the planets having $2 \leq R_{pl} \leq 4$ $R_\oplus$ and $P \leq 50$ days. On the other hand, preliminary results from the HARPS planet search (Lovis et al., 2009; Mayor et al., 2009; Udry, 2010) indicate that 30 - 50% of Sun-like stars host sub-Neptune mass planets within 50-day orbits [1] — a planet frequency that is substantially higher than the *Kepler* occurrence rate. Although these two occurrence rates do provide somewhat different information, as discussed in §2.5.1, the following order-of-magnitude argument readily gives a sense for the apparent discrepancy in terms of the total number of planets that *Kepler* would have detected in its first four months of data.

[1] After our manuscript was submitted, Mayor et al. (2011) published more detailed statistical results from the HARPS survey. Their data do not appear to be in conflict with the analysis of this investigation.

Given a 40% occurrence rate and $\sim$ 150,000 *Kepler* target stars, there are 60,000 potentially detectable planets in *Kepler*'s field of view, assuming that each host star harbors only one planet. Not all of these planets will transit, however, as the required star-planet-observer alignment is fairly improbable given random inclinations along the line of sight. For planets in orbits of 50 days or less, this geometrical transit probability works out to be $1 - 15\%$; taking a 5% transit probability (10-day orbit) as a benchmark, the number of sub-Neptune-mass planets that *Kepler* would have been able to detect is thus approximately 3,000. If we map the *Kepler* planet candidate radii to mass via the simple relation $M/M_{\oplus} = (R/R_{\oplus})^{2.06}$ (Lissauer et al., 2011b), then we see that about 900 of *Kepler*'s planet candidates fall in the $M < \mathrm{M}_{Nep}$ range. Thus, the HARPS occurrence rate appears to overestimate the number of planets that *Kepler* would have detected by a factor of 3.

Order-of-magnitude arguments can be misleading, however, so in this chapter we take care to fully account for details of the RV-transit comparison that may affect this result, including factors such as the enhanced geometrical transit probability of elliptical orbits, the shallower transits of more inclined orbits due to stellar limb darkening, target star selection biases, and *Kepler*'s detection incompleteness. In conducting the comparison, we primarily investigate the effect that two different M-Rs have on the total number of planets that could have been detected by *Kepler* after four months, assuming the HARPS occurrence rate. In doing so, we identify the parameter values for two different mass-to-radius relationships that produce agreement between the total numbers of short-period low-mass planets observed by both RV and transit searches, and

we examine how these M-Rs influence the interpretation of RV and transit occurrence rates.

The layout of this chapter is as follows. In §5.2 we briefly summarize the *Kepler* data set and discuss the use of planet candidates instead of confirmed planets for our analysis. In §2.3 we describe our simulations and the *Kepler* planet candidate sample we use in our comparison. In §2.4 we present our results on the total number of planets that *Kepler* would have been able to detect given the HARPS occurrence rate, and in §5.5 we discuss the implications of these results for our current understanding of exoplanet populations.

2.2 The *Kepler* Data Set

On February 1, 2011 *Kepler* released its second quarter (Q2) data, which was soon followed by the announcement of 1,235 transiting planet candidates (Borucki et al., 2011). It is necessary to note, however, that the vast majority of these planets are unconfirmed and thus maintain "planet candidate" status. The current consensus is that these candidates can be catalogued as true planets only if they exhibit transit timing variations or are detected through the radial velocity method, as other astrophysical events such as binary blends with background stars, eclipsing hierarchical triples with small separations, and certain types of stellar variability can mimic planetary transits (Gautier et al., 2010; Morton & Johnson, 2011b). Unfortunately, however, the majority of *Kepler*'s target stars have $V > 11$ and thus are faint for the purpose of Doppler follow-up, making these additional RV measurements expensive and leaving the vast

majority of the *Kepler* candidates unconfirmed.

To compensate for these observational limitations, the *Kepler* team has developed an extensive vetting process to eliminate as many of these false planetary transit signatures as possible (Gautier et al., 2010; Borucki et al., 2011). Inevitably, however, a small but non-negligible fraction of false positives are expected to persist in the list of planet candidates. Borucki et al. (2011) estimates that this false positive fraction is as high as 20%, while a detailed Bayesian analysis conducted by Morton & Johnson (2011b) finds that the transit depth-independent false alarm probability is $< 5\%$ over the entire field of view, given stars with *Kepler* magnitude Kp ≤ 16, a 30-50% planet occurrence prior, and the assumption that follow-up astrometry can identify binaries at any Kp with separations $> 2''$. When this last assumption is relaxed, as is necessary for the planet candidates reported by Borucki et al. (2011), the false alarm probability increases with decreasing transit depth and can exceed 30% for Kp > 15. In proceeding with our statistical analysis of the *Kepler* planet candidate population, we thus bear in mind that the total number of true *Kepler* planets is likely $\sim 5\%$ lower than that reported by Borucki et al. (2011) for Neptune-sized planets and as much as 30% lower for Earth-sized planets around dim stars.

2.3 The Transit-RV Comparison: Methods

Comparing the HARPS occurrence rate with *Kepler*'s planet candidates involves several steps. First, we require that the aggregate properties of our initial planet population are consistent with the cumulative characteristics of the low-mass population

detected by the HARPS survey (§2.3.1). To compare these planets to *Kepler*'s public data set, we map our initial distribution of planet masses to radii via a population-wide mass-to-radius relationship (M-R; §2.3.2). Each simulated planet is subsequently matched to a *Kepler* target star (§2.3.3) and its light curve is computed based on analytic transit formulae (Mandel & Agol, 2002). We then apply *Kepler*'s detection criteria (Batalha et al., 2010b) to assess whether or not that planet would have been detected by the end of the second quarter (§2.3.4). Finally, for a range of parameter sets, we tabulate the total number of planets that *Kepler* would detect in its first four months of data (N_{detect}) when the underlying planet population conforms to the HARPS occurrence rate, and compare this number to the total number of analogous planets that *Kepler* actually does detect.

2.3.1 Simulations of the Radial Velocity Population

Other than stating that 30 - 50% of Sun-like stars host sub-Neptune-mass planets with $P \leq 50$ days, the HARPS overall occurrence rate does not address specific details of the planetary mass-period distribution. Accordingly, we must select a general, easily parameterized distribution that is able to recover the HARPS overall occurrence rate. Power laws meet these criteria, so we follow common practice (Cumming et al., 2008; Howard et al., 2010, 2012; Youdin, 2011) and adopt:

$$N(M)dM = N_{tot}C_M M^{\alpha} dM, \tag{2.1}$$

where $N(M)dM$ is the number of planets that have a mass between M and $M + dM$, N_{tot} is the total number of planets in the sample, C_M is a normalization constant, and α is the mass power law index. Similarly, we take for the period distribution:

$$N(P)dP = N_{tot}C_P P^{\beta}dP. \tag{2.2}$$

We use the HARPS overall occurrence rate to determine N_{tot}, C_M, and C_P for our simulated populations. N_{tot} is simply the planet occurrence rate times the total number of stars that *Kepler* is observing, assuming that each star which harbors a planet harbors no more than that one planet—the bare minimum suggested by the prediction. Given that *Kepler* observed over 110,000 G and K dwarfs during its second quarter (Q2) of data (Kepler Data Release 7, Multimission Archive at STScI), this leads to $N_{tot} \sim 55000$ for a 50% occurrence rate. C_M and C_P are determined by setting minimum and maximum values for mass and period in our simulations. The maximum period of 50 days is explicitly given by the stated HARPS occurrence rate, as are the limits on planet mass if we define a sub-Neptune planet to have $1 \leq M \leq 17$ $M_{\oplus} \sim M_{Nep}$. It is important to emphasize that we are only considering low planetary masses here; Jupiter-mass planets are not considered in our simulations.

The minimum value on period, while not expressly indicated in the HARPS low-mass occurrence rate, can be reasonably chosen from existing trends. Both the census of *Kepler* planet candidates (Borucki et al., 2011) and the population of planets discovered through the radial velocity method (Wright et al., 2011) suggest that there is a dearth of planets with $P < 2$ days that is not due to the selection biases of the

different detection methods. Howard et al. (2012) fit a power-law distribution with an exponential cutoff at short periods to the *Kepler* planet candidates and found that the transitional period varies from 2 to 7 days for planets with $2 \leq R \leq 6$ $R_{\oplus}$. To simplify our input distributions, we ignore the exponential cutoff and set $P_{min} = 2$ days, keeping in mind that any deviation from a power law for $2 \leq P \leq 7$ days may impact our ability to fit *Kepler*'s observed distribution.

A rigorous interpretation of the HARPS statistic would include the unknown $\sin(i)$ factor on the observed masses. We note, however, that the distribution of inclinations for the observed radial velocity planets is poorly understood and that spherical isotropy cannot be assumed due to the detection biases inherent in the radial velocity technique. Although some insight may be gleaned from statistical analysis such as that in Ho & Turner (2011) or from the few planets which exhibit the Rossiter-McLaughlin Effect (Schlaufman, 2010), in this analysis we take our mass limits as the bounds on the true mass of our simulated planets, effectively ignoring any refinements stemming from the $\sin(i)$ effect.

Simulation Parameters

To account for the ambiguity in the RV mass and period distributions, we require that the power-law indices α and β serve as free parameters in our simulations: we allow α to vary from -2.5 to 0 and β to vary from -0.5 to 0.5, both in increments of 0.1. We model eccentricity, e, longitude of periastron, ω, and inclination, i, as uniform distributions, randomly drawing e from $0 \leq e \leq 0.2$, ω from $0 \leq \omega < 2\pi$, and i from an isotropic sphere. Taken with P, these orbital elements serve to determine which

planets transit, given their geometrical transit probability. We choose to include non-zero eccentricities because elliptical orbits can enhance the probability of a transit, but we set the upper bound at $e = 0.2$ with the expectation that many short-period planets will have experienced a significant degree of tidal circularization. This bound is broadly consistent with the observed eccentricity distribution of confirmed planets in our mass and period range, which shows that a vast majority ($\sim 80\%$) of low-mass planets with $P < 50$ days have $e \lesssim 0.2$ (Wright et al., 2011).

Two more free parameters are introduced for the second M-R we consider in this chapter (§2.3.2), as we allow the fraction of rocky planets in the population to vary as a linear function of mass. These fractions are then used to randomly assign each planet either a gaseous or a rocky composition. In addition, we randomly allocate each planet to a *Kepler* target star, as discussed in §2.3.3.

2.3.2 Population-wide Mass-to-Radius Relationships

A crucial consideration for the transit-RV comparison is the population-wide M-R used to map an RV planet's mass to a transiting planet's radius. Howard et al. (2012) have shown that applying one bulk density to an entire planet population fails to match the *Kepler* candidates, so we begin our investigation with more flexible and physically motivated M-Rs, while taking care to minimize the number of degrees of freedom. In particular, we consider two population-wide M-Rs in this chapter: a power-law fit to measured planetary masses and radii, and a multi-valued parameterization that relaxes the single-valued assumption involved in fitting a power law to data.

Single-Valued M-R

Lissauer et al. (2011b) use the following power-law fit to Earth and Saturn as the mass-radius relation for *Kepler*'s planet candidates:

$$\frac{M}{M_\oplus} = \left(\frac{R}{R_\oplus}\right)^{2.06}, \tag{2.3}$$

which tacitly assumes that extrasolar planets resemble those in our Solar System. Experience has shown that such an approach requires caution, so as a check we derive a comparable M-R for the five transiting extrasolar planets with $1 \leq M \leq 17$ $M_\oplus$ and $2 \leq P \leq 50$ days (Wright et al., 2011): Kepler-11 b, c, d, e, and f (Lissauer et al., 2011a). Including the error on both mass and radius, we employ angular bisector least squares regression to find the following best fit:

$$\frac{R}{R_\oplus} = 0.95^{+0.66}_{-0.26}\left(\frac{M}{M_\oplus}\right)^{0.66\pm0.17}, \tag{2.4}$$

which is consistent with the inverse of Equation 2.3 within 1σ. We note, however, that the M-R computed directly from the dually-detected, low-mass extrasolar planets has large error bars and is poorly constrained, so we proceed cautiously with $R/R_\oplus = (M/M_\oplus)^{0.48}$ as implied by Equation 2.3.

Multi-Valued M-R

While the Lissauer et al. (2011b) M-R implicitly incorporates compositional variation from planet to planet, it does not allow for the possibility of a multi-valued M-

R. This is potentially a severe shortcoming, as a more complex M-R has appeared to be necessary from the outset of observational constraints on low-mass planet compositions: the first two such planets with measured radii and masses, CoRoT-7 b (Queloz et al., 2009; Léger et al., 2009) and GJ 1214 b (Charbonneau et al., 2009), yielded very different densities (6 g cm^{-3} and 2 g cm^{-3} respectively), despite having similar masses (4.9 $M_\oplus$ and 6.5 $M_\oplus$).

With this observational evidence in mind, we believe that the key to reconciling the *Kepler* and HARPS occurrence rates may be a multi-valued low-mass M-R, where more than one planetary radius is allowed at each planetary mass. Our parameterization assumes that the simulated planets can have either a significant gaseous composition (Neptune analogs; an extension of the gas giants to lower masses) or a rocky composition (Earth analogs; an extension of the terrestrial planets to higher masses), and that the admixture of these two compositions varies as a linear function of mass for $1 \leq M \leq 17$ $M_\oplus$. This admixture is quantified by the fraction of rocky planets in the population, $f_{rocky}(M)$: if, for example, $f_{rocky}(1) = 1.0$ and $f_{rocky}(17) = 0.0$, then $f_{rocky}(8) = 0.5$, meaning that all 1 $M_\oplus$ planets would be rocky, all 17 $M_\oplus$ planets would be gaseous, and the 8 $M_\oplus$ planets would be evenly divided between the two compositions. In our simulations we allow $f_{rocky}(1)$ and $f_{rocky}(17)$ to vary between 0 and 1 in increments of 0.1, giving two more free parameters in our simulations (we also allow the same α that is present for the single-valued M-R to vary between -2.0 and -0.6 in increments of 0.1, and the same β to vary between 0.2 and -0.2 in increments of 0.1).

For the Earth analogs in this multi-valued M-R we use the Solar System's

terrestrial planet population-wide mass-to-radius relationship: $R/R_\oplus = (M/M_\oplus)^{0.33}$. We emphasize that this rocky M-R is not just a re-expression of the individual mass-to-radius relationship for a constant-density sphere; instead, this population-wide M-R was derived by fitting a power law to all of the Solar System's inner planets, much like the $R \propto M^{0.48}$ relationship was derived above.

For the Neptune analogs we use the M-R curves calculated by Rogers et al. (2011). These authors model the structure of low-mass planets with substantial gaseous envelopes by invoking a core accretion formation history and then self-consistently incorporating the effect of planetary equilibrium temperature, T_{eq}, across the range of orbital periods and stellar fluxes that we consider here. They find, however, that the M-R curves of constant gaseous envelope mass fraction, M_{env}, are remarkably insensitive to planet mass above ~ 7 $M_\oplus$. Because no single M_{env} provides the dynamic range needed to explain the diversity of radii that *Kepler* observes, we must allow for variation in envelope fraction to construct a population-wide M-R that can reasonably reproduce the observed radius range. Noting that the M-R curves are roughly equally spaced in R by approximately logarithmic bins in M_{env}, we randomly choose an envelope mass fraction from a log-uniform distribution between 10^{-5} and 10^{-1}. Finally, using Figure 4 of Rogers et al. (2011), we interpolate our simulated planets' radii as a function of M, M_{env}, and T_{eq}.

As Rogers et al. (2011) illustrates, varying M_{env} allows planets with masses as small as 2 $M_\oplus$ to have a radius as large as 7 $R_\oplus$, which enables planets less massive than Neptune to fall within the $2 \leq R \leq 6$ $R_\oplus$ range that *Kepler* has found to be well

populated (Borucki et al., 2011; Howard et al., 2012). However, these relatively low-mass, large-radius planets are particularly susceptible to atmospheric mass loss, and so these planets may not actually be able to hold onto their gaseous envelopes, depending on the amount of irradiation they receive from their host star. Following the discussion in Rogers et al. (2011), we incorporate the possibility of mass loss in our population-wide M-R via the following timescale argument.

As illustrated by Lammer et al. (2003), one must consider the effects of X-ray and extreme ultraviolet (XUV) irradiation on a planet's thermal structure in order to realistically treat atmospheric mass loss. In the regime where the amount of energy incident on the planet determines the degree of atmospheric escape, this mass loss is parameterized by (Lecavelier Des Etangs, 2007; Valencia et al., 2010; Rogers et al., 2011)

$$\dot{M} = -\frac{\epsilon \pi F_{XUV} R_{XUV}^2 R_p}{G M_p K_{tide}}, \tag{2.5}$$

where F_{XUV} is the XUV flux incident on the planet from the host star; ϵ is the fraction of incident XUV energy that is actually absorbed by the atmospheric particles; R_{XUV} is the planet radius at which the XUV flux is absorbed; R_p is the radius of the planet as calculated from planetary interior structure models; M_p is the mass of the planet; and K_{tide} is a tidal correction factor of order unity for planets with $R \lesssim R_{Nep}$ and $P > 2$ days. Unfortunately, ϵ is largely unknown, so at best Equation 2.5 provides an order-of-magnitude estimate for $\dot{M}$. We follow Rogers et al. (2011) in setting $\epsilon = 0.1$ and $F_{XUV} = F_{XUV,\odot} = 4.6 \times 10^{-3}$ W m^{-2} (Ribas et al., 2005); we scale F_{XUV} by the equilibrium temperature of the planet, which depends on the radius of the host

star, the effective temperature of the host star, and the semimajor axis of the planet's orbit. From the mass loss timescales plotted by Rogers et al. (2011) we estimate that $R_{XUV}^2 \sim 10R_p^2$ for these short-period low-mass planets. With $\dot{M}$ thus determined, the atmospheric mass loss timescale is

$$t_{loss} = -\frac{M_{envelope}}{\dot{M}}. \tag{2.6}$$

If $t_{loss} < 1$ Gyr, we consider the planet to have completely lost its gaseous envelope, and we take the radius of the planet to be the radius of its 50% rock, 50% ice core (Fortney et al., 2007).

2.3.3 Star Selection

Once we apply an M-R to the simulated RV population, we randomly allocate planets to specific *Kepler* target stars. This one-to-one matching allows us to sidestep the concern that the selection biases exhibited by different detection methods will significantly influence computed planet occurrence rates (Howard et al., 2012), and we can directly compare our simulated population with *Kepler*'s planet candidates. Accordingly, we adopt the list of 165,000 long-cadence Q2 *Kepler* target stars to initiate our star selection. We begin by extracting the photometrically-derived effective temperature, T_{eff}, the surface gravity, $\log(g)$, the radius, R_{star}, and the *Kepler*-bandpass apparent magnitude, Kp, from the each star's Q2 FITS header. These data originate from the Kepler Input Catalog (KIC; Kepler Mission Team, 2009), which has known errors of $\pm$200 K on T_{eff} and $\pm$0.4 dex on $\log(g)$ (Brown et al., 2011). Because these

two parameters are used to calculate R_{star}, the errors on the planet candidates' radii can be significant; in §5.5 we discuss the possible effect of these errors on our results.

In their analysis of *Kepler*'s planet candidates, Howard et al. (2012) compute *Kepler*'s observed occurrence rates from a heavily vetted list of target stars, whose total noise in one quarter of data enables detection of a $R \geq 2$ $R_\oplus$ planet with SNR ≥ 10. This approach prompts them to drop all stars with $Kp \geq 15$ and all planets with $R < 2$ $R_\oplus$ due to concerns about sample incompleteness. By contrast, our approach retains the entire *Kepler* target star sample, with only the $\log(g)$ cut discussed below. Because we individually simulate each planet's light curve to accurately determine its detectability (§2.3.4) and then ask how many planets *Kepler* would have seen in its first four months of data if the HARPS occurrence rate is true (§2.4), we naturally account for the detection incompleteness in *Kepler*'s first four months of data; this incompleteness is displayed graphically in Figure 2.1 as the smallest-radius planet that each star could have detected by the end of Q2. Thus, our simulation procedure permits us to include dimmer stars and smaller planets with radii down to 1 $R_\oplus$, which allows us to draw conclusions with a larger sample size and improved statistics.

The only severe cut we make to the 165,000 available Q2 target stars is in $\log(g)$. We restrict potential planet-hosting stars to those with $\log(g) > 4.0$ to minimize contamination from subgiants, as the KIC's surface gravities are poorly contrained above $T_{eff} \sim 5400$ K (Brown et al., 2011). The resulting list consists of 131,000 stars (Figure 2.1), the vast majority ($> 110,000$) of which are G and K dwarfs. Nonetheless, a small proportion of subgiants and giants, whose radii may be underestimated in the KIC by

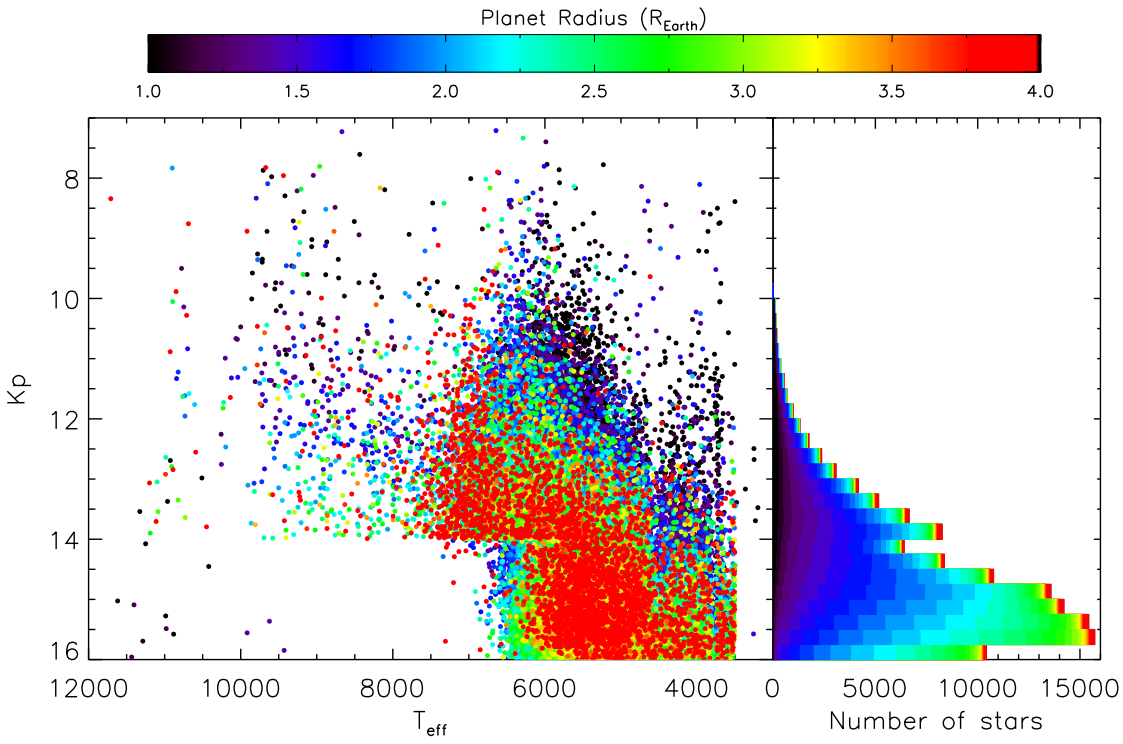

Figure 2.1: Left: apparent magnitude, Kp, and effective temperature, T_{eff}, from the Kepler Input Catalog (KIC) for the *Kepler* target stars included in our simulations (§2.3.3). All of these stars have KIC $\log(g) > 4.0$. Right: number of $\log(g) > 4.0$ *Kepler* target stars in each apparent magnitude bin. The color represents the smallest planet around each target star that *Kepler* could have detected in its first four months of data, assuming an orbit with $P = 20$ days and $e = 0$. With the same orbital parameters for each size planet, this minimum detectable radius is thus determined by the radius of the star, R_{star}, and by the star's total photometric noise on a three-hour timescale, $CDPP_3$ (§2.3.4). In the scatterplot, note the general trend of minimum detectable radius with both Kp and T_{eff}, which correlate with $CDPP_3$ and R_{star}, respectively. The histogram to the right more clearly illustrates the trend of increasing minimum detectable radius with increasing Kp (due to increasing $CDPP_3$). However, it is important to note that there do exist dim target stars around which *Kepler* could have already detected a 1 to 1.5 $R_\oplus$ planet. This is a result of the trend of decreasing minimum radius with decreasing R_{star} and the fact that low-mass stars exist in every Kp bin.

as much as a factor of 2 (Brown et al., 2011), likely remains in our target star sample.

Without knowledge about the degree of subgiant contamination, we cannot accurately

account for their statistical effect in our results, although we expect that this effect will be very small based on the low numbers of possibly misclassified evolved stars found by Basri et al. (2011).

2.3.4 Detectability of Simulated Planets

To pinpoint the simulated planets that *Kepler* would have identified as planet candidates after four months of data collection, we first compute analytic light curves (Mandel & Agol, 2002) for the simulated planets that transit according to their geometric transit probability (Seagroves et al., 2003). These light curves incorporate the planets' eccentricity and inclination as well as the *Kepler*-bandpass limb darkening coefficients that are calculated by Claret & Bloemen (2011) for a large range of stellar effective temperatures, surface gravities, and metallicities. Using a 30-minute cadence over 132 days to match *Kepler*'s long-cadence Q0 - Q2 datasets, we determine the transit depth, duration, and the total number of transit events directly from the simulated light curves.

As described in Batalha et al. (2010b), *Kepler*'s detectability criterion is set such that less than one false positive planet detection over its 3.5 year mission would be expected to result from purely statistical fluctuations in stellar photon counts. This requirement gives a 7.1σ threshold for a transit's statistical significance when the light curve is folded and binned. The detectability of a planet therefore depends on both R_{pl}/R_{star} and a number of stellar parameters and instrumental properties which affect the total noise (Batalha et al., 2010b; Jenkins et al., 2010a). These systematic errors are difficult to assess without intimate knowledge of *Kepler*'s performance, so we use the

noise calculated directly by the *Kepler* data reduction pipeline, the Combined Differential Photometric Precision (CDPP, Christiansen et al. 2012), to reproduce as accurately as possible the planet population that *Kepler* could have identified by the end of Q2.

Defined as the root mean square of stellar photometric noise on transit timescales, the CDPP provides the most accurate estimate of the noise from each target star that would interfere with a transiting planet's detectability. A wavelet-based, adaptive matched filter is applied to the corrected *Kepler* light curves in the Transiting Planet Search section of the Science Processing Pipeline (Jenkins et al., 2010c) to produce 3-hour, 6-hour, and 12-hour CDPP estimates, which are then used to calculate the statistical significance of a possible transit event. Incorporating *Kepler*'s own noise metric in our simulations automatically folds in its detection biases and accounts for sample incompleteness below 2 $R_{\oplus}$; therefore, we can extend our analysis down to Earth-sized planets without reservations about hidden selection effects.

Our simulations only consider planets with $2 \leq P \leq 50$ days, so the 3-hour CDPP estimate is the most relevant for our purposes. Matching each planet to a *Kepler* target star also matches it to a CDPP value, so we scale this noise estimate by the transit duration and the total number of transit events observed during Q0 - Q2 (Batalha et al., 2010b; Howard et al., 2012). Our detectability criterion therefore becomes:

$$SNR = \frac{\delta \sqrt{N_{tr} \frac{N_{dur}}{6}}}{CDPP_3} > 7.1, \tag{2.7}$$

where $\delta \propto (R_{pl}/R_{star})^2$ is the maximum transit depth (in ppm) identified from the analytic light curves, $CDPP_3$ is the Q2 3-hour Combined Differential Photometric Pre-

cision (in ppm) associated with the planet's host star, N_{tr} is the number of observed transits in four months, and N_{dur} is the number of data points acquired per transit on a 30-minute cadence. We note that δ is proportional but not equal to $(R_{pl}/R_{star})^2$ because we include a range of possible transit-producing inclinations and self-consistently incorporate the effect of limb darkening based on the host star's T_{eff} and $\log(g)$.

Figure 2.1 illustrates our detectability criterion graphically, with the color scale showing the smallest planet for each $\log(g) > 4.0$ target star that *Kepler* could have detected after four months of data collection, assuming an orbit with $P = 20$ days and $e = 0$. As expected, this minimum detectable R_{pl} trends with both Kp and T_{eff}, which correlate with CDPP and R_{star}, respectively. When the orbit is not held constant, an individual planet's detectability is also determined by its orbital period, as given by N_{tr} in Equation 2.7.

2.3.5 Comparison with Total Number of True *Kepler* Planets

To fairly conduct the transit-RV comparison, we also need to filter the list of 1,235 *Kepler* planet candidates to match the limits we impose on our simulated population. Accordingly, we retain only those candidates with $1 \leq R < 4$ $R_\oplus \sim R_{Nep}$ and $2 \leq P \leq 50$ days orbiting stars with $Kp \leq 16$. We also impose a cut on the candidates in multiple-planet systems, including only the first planet listed by the *Kepler* Science Processing Pipeline in this mass and period range; in most cases, this is the planet labeled ".01". This cut conforms to our assumption of one planet per host star and reduces the total number of *Kepler* planets in our radius and period range from 797 to 631, a difference of 166 ($\sim 20\%$).

Of course, we must also account for the probability of false positives among *Kepler*'s planet candidates. As stated in §5.2, the total number of true *Kepler* planets is likely $\sim$ 5% lower than that reported by Borucki et al. (2011) for Neptune-sized planets and may be as much as 30% lower for Earth-sized planets around the dimmest stars (Morton & Johnson, 2011b). Applying the 5% false positive rate across the board reduces the total number of actual planets in our filtered list to 599; when we take into account the effect of planet size and stellar apparent magnitude as suggested by Figure 8 of Morton & Johnson (2011b), the 339 planets with $R < 3$ $R_\oplus$ and Kp > 14 contain about 50 additional false positives. Thus, in the worst case scenario we should compare the total number of detectable simulated planets to $\sim$ 550 rather than 631. Additional studies such as Désert et al. (2015), however, suggest that the *Kepler* false positive rate is even lower than 5%, and thus $\sim$ 600 - 630 is the appropriate number to compare to. Considering these results, we adopt 631 for straightforwardness. In doing so, we are also assuming that the completeness of the *Kepler* Science Processing Pipeline (Jenkins et al., 2010b) is high, meaning that *Kepler* has detected nearly all of the planets it should have been able to detect.

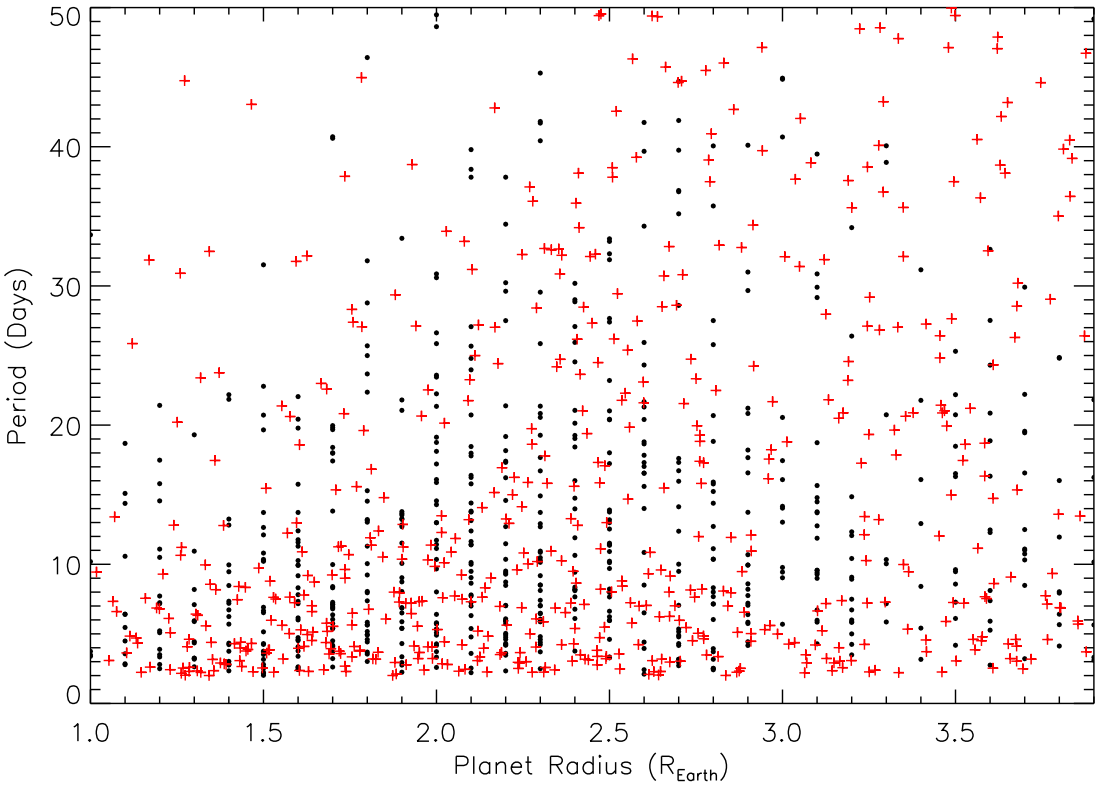

Figure 2.2: Period vs. radius for a single realization of the simulated planet population produced by the $R = M^{0.48}$ mass-to-radius relationship with $\alpha = -1.7$, $\beta = 0.0$, and a 40% overall occurrence rate. The *Kepler* planet candidates are marked with the black circles, and the detectable simulated planets are marked with the red plus signs. There are 562 simulated planets in this realization; the total number of *Kepler* planet candidates here is 631.

2.4 The Transit-RV Comparison: Results

2.4.1 Single-Valued M-R

The above procedure gives us the period-radius distribution that *Kepler* would detect in its first four months of data when the underlying planet population conforms to the HARPS occurrence rate. An example of our simulations' output is given in Figure 2.2, which represents one realization of the single-valued mass-to-radius relationship

(§2.3.2) computed at a 40% overall occurrence rate with the parameter values $\alpha = -1.7$ and $\beta = 0.0$ (corresponding to the mass power law index and the period power law index, respectively: §2.3.1).

Our lack of detailed knowledge about the HARPS data set necessitates that we allow for some freedom in the population's initial mass and period distributions. Incorporating this freedom has a significant effect on the total number of planets that *Kepler* should have been able to detect in its first four months of data, even as we hold the HARPS overall occurrence rate constant at 40%. This effect is quantified in Figure 2.3, which summarizes the result of 100 realizations of the single-valued M-R for all sets of parameter values that we consider ($-2.5 \leq \alpha \leq -1.0$ and $-0.5 \leq \beta \leq 0.5$). The color denotes the total number of detectable (§2.3.4) simulated planets (N_{detect}) with $1 \leq R \leq 3.9$ $R_\oplus = (17$ $M_\oplus)^{0.48}$ and $2 \leq P \leq 50$ days, averaged over all N=100 realizations with a standard deviation of roughly 30; the parameter sets with a red color roughly produce the total number of analogous *Kepler* planet candidates in our filtered list (631).

In an effort to identify which of these parameter sets best fit *Kepler*'s planet candidates, we employed the two-sample two-dimensional Kolmogorov-Smirnov (2-D K-S) test (Fasano & Franceschini, 1987). We chose to use the 2-D K-S statistic because it avoids binning data, unlike the more common χ^2 test, and thus maximally preserves information contained in the planets' radius-period distributions. When we perform parametric Monte Carlo bootstrap resampling of this statistic to compute confidence levels, we find that all of these parameter sets are ruled out at the $P < 0.001$ level after

1000 realizations, with the closest parameter set, $\alpha = -1.7$ and $\beta = 0.0$, producing a K-S statistic that is on average 10.8 standard deviations from the mean of its bootstrapped K-S distribution. Thus, the single-valued M-R with our simplified parameterization of the planet population is insufficient to reproduce the details of the *Kepler* planet candidate period-radius distributions. Given that our primary focus is on the total number of planets detectable by *Kepler*, however, we note that there is a locus of parameter sets between $-2.3 \leq \alpha \leq -1.0$ and $-0.5 \leq \beta \leq 0.5$ which can produce N_{detect} consistent with the total number of analogous *Kepler* planet candidates. We also find that N_{detect} varies linearly with the occurrence rate, and so this locus will shift up and down in parameter space accordingly: a 50% occurrence rate, for example, would produce consistent N_{detect} for $-2.5 \leq \alpha \leq -1.3$ and $-0.5 \leq \beta \leq 0.5$ (the purple and maroon regions in Figure 2.3).

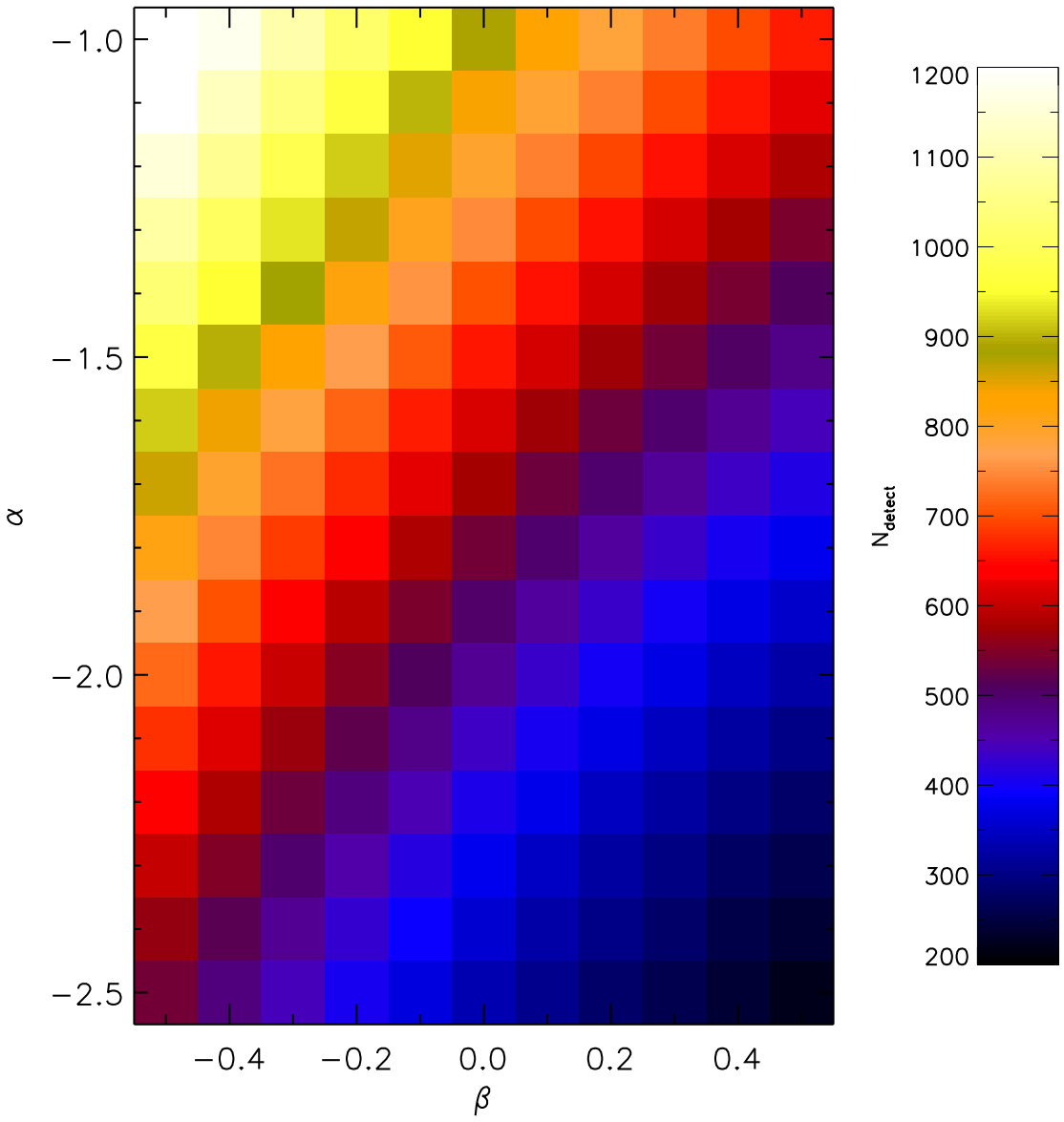

Figure 2.3: The total number of detectable (§2.3.4) simulated planets (N_{detect}) produced by the $R = M^{0.48}$ mass-to-radius relationship (§2.3.2) for a 40% overall occurrence rate in the $2 \leq P \leq 50$ days and $1 \leq R < 4$ $R_{\oplus}$ range; this total number is averaged over 100 of the realizations illustrated in Figure 2.2, for each parameter set, and has a standard deviation of roughly 30. The axes denote the period power law index (β) and mass power law index (α), which serve as free parameters in our simulations (§2.3.1). The parameter sets with a darker red color roughly produce the total number of analogous *Kepler* planet candidates in our filtered list (631).

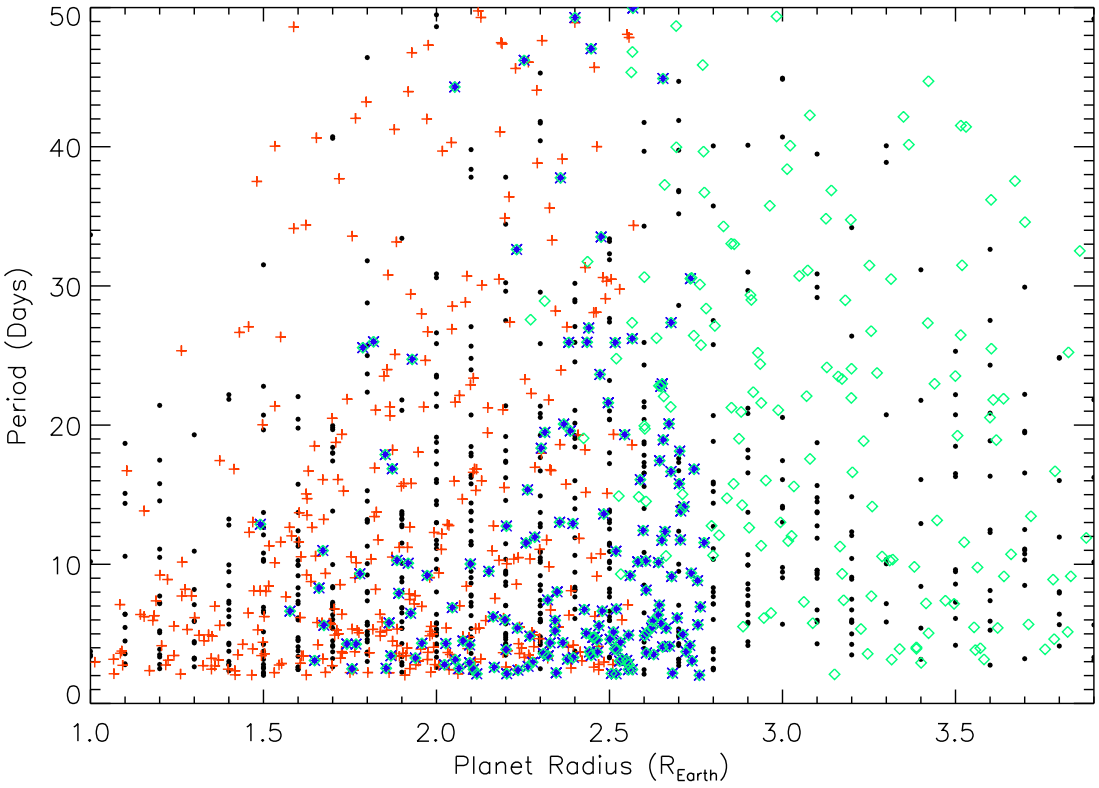

Figure 2.4: Period vs. radius for a single realization of the simulated planet population produced by the multi-valued mass-to-radius relationship with $\alpha = -0.9$, $\beta = 0.0$, $f_{rocky}(1) = 1.0$, $f_{rocky}(17) = 0.2$, and a 40% overall occurrence rate. The black circles denote the *Kepler* planet candidates; the red plus signs denote the detectable simulated planets with a rocky composition; the green diamonds denote the detectable simulated planets with a gaseous composition; and the blue asterisks denote the detectable simulated planets with a half-rock, half-ice composition, which could be produced by significant mass loss from the gaseous planets. There are 666 simulated planets in this realization; the total number of *Kepler* planet candidates here is 631.

2.4.2 Multi-Valued M-R

Results analogous to those in §2.4.1 but for our multi-valued M-R are presented in Figures 2.4 and 2.5. An example of our simulations' output is given in Figure 2.4, which represents one realization of the multi-valued mass-to-radius relationship (§2.3.2)

computed at a 40% overall occurrence rate with the parameter values $\alpha = -0.9$, $\beta = 0.0$, $f_{rocky}(1) = 1.0$, and $f_{rocky}(17) = 0.2$ (corresponding to the mass power law index, the period power law index, the fraction of all 1 $M_\oplus$ planets that have a rocky composition, and the fraction of all 17 $M_\oplus$ planets with a rocky composition, respectively: §2.3.2).

As with the single-value M-R, N_{detect} depends sensitively on the free parameters in our simulations. The result of 100 realizations of this multi-valued M-R is summarized in Figure 2.5 for a subset of the parameter values that we consider. Again, the color shows the total number of detectable simulated planets (N_{detect}) averaged over 100 realizations for a 40% overall occurrence rate, with a standard deviation of roughly 30; the parameter sets with a red color roughly produce the total number of analogous *Kepler* planet candidates in our filtered list (631).

When we apply the 2-D K-S test to these populations, we again find that all of these parameter sets are ruled out at the $P < 0.001$ level after 1000 realizations, with the closest parameter set, $\alpha = -0.9$, $\beta = 0.0$, $f_{rocky}(1) = 1.0$, $f_{rocky}(17) = 0.2$, producing a K-S statistic that is on average 7.7 standard deviations from the mean of its bootstrapped K-S distribution. Thus, the multi-valued M-R with our simplified parameterization of the planet population is still insufficient to reproduce the details of the *Kepler* planet candidate period-radius distributions; the implications of this are discussed in §5.5. As we are primarily concerned with the total number of planets that *Kepler* would have observed, however, it suffices to note that there are a number of parameter sets at realistic values of $f_{rocky}(1)$, and $f_{rocky}(17)$ which can produce N_{detect} consistent with the total number of analogous *Kepler* planet candidates.

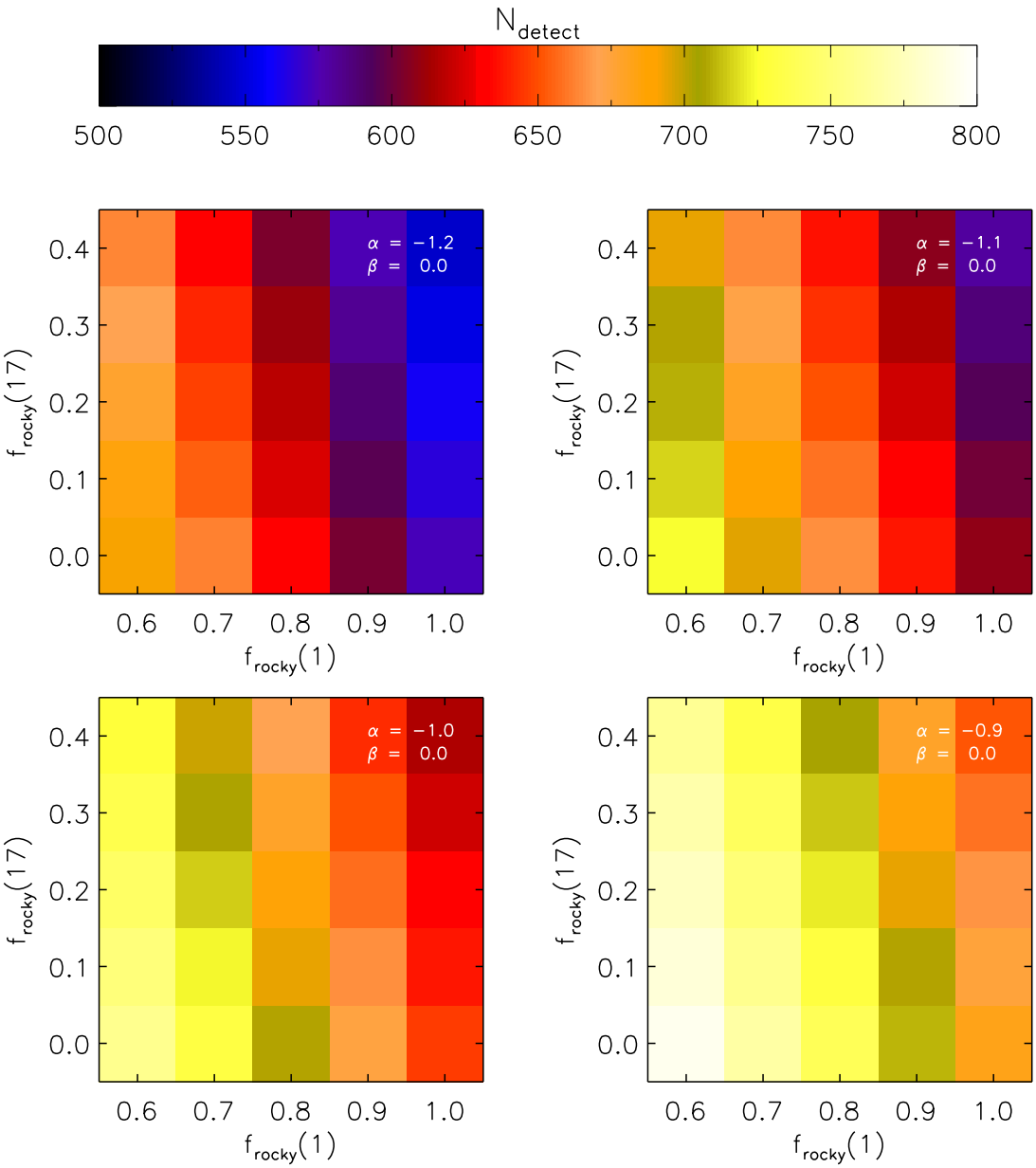

Figure 2.5: The total number of detectable (§2.3.4) simulated planets (N_{detect}) with $2 \leq P \leq 50$ days and $1 \leq R < 4$ $R_\oplus$ produced by the multi-valued mass-to-radius relationship (§2.3.2) for a 40% overall occurrence rate; this total number is averaged over 100 of the realizations illustrated in Figure 2.4, for each parameter set, and has a standard deviation of roughly 30. The axes denote the fraction of all 1 $M_\oplus$ planets in the simulated planet population that have a rocky composition, $f_{rocky}(1)$, and the fraction of all 17 $M_\oplus$ planets that have a rocky composition, $f_{rocky}(17)$; each panel corresponds to a different value of the mass power law index (α) at a constant period power law index of $\beta = 0.0$. The parameter sets with a red color roughly produce the total number of analogous *Kepler* planet candidates in our filtered list (631).

2.5 Discussion

One of the most important questions driving the search for extrasolar planets is the frequency with which planets of different sizes, masses, and semi-major axes occur in our Galaxy. The frequency of super-Earths and sub-Neptunes is particularly interesting for the hints it gives on the frequency of Earth-like planets. Interestingly, the two surveys that have detected the most super-Earths/sub-Neptunes to date appear to disagree on the overall occurrence rate of these kinds of planets: the first four months of data from the *Kepler Mission* yielded 0.130 ± 0.008 (Howard et al., 2012) or 0.19 (Youdin, 2011) planets with $2 \leq R_{pl} \leq 4$ $R_\oplus$ and $P \leq 50$ days per solar-type star, while preliminary results from the HARPS planet search (Lovis et al., 2009; Mayor et al., 2009; Udry, 2010) indicated that 30 - 50% of Sun-like stars host sub-Neptune mass planets within 50-day orbits. While these two occurrence rates are not immediately comparable to each other (see §2.5.1), the correction needed to make their definitions coincide actually causes the occurrence rates to become even more discrepant. With this in mind, we undertook a Monte Carlo study to investigate physically intuitive explanations for this apparent disagreement, in terms of the total number of planets that *Kepler* would have detected in its first four months of data given the HARPS occurrence rate (N_{detect}).

To this end, we note that *Kepler* and HARPS measure fundamentally different properties of the detected planets: transit surveys measure radius, while RV surveys measure mass with an unknown inclination factor. Because super-Earth/sub-Neptune planets lie at the detection threshold of both surveys, it seems feasible that these two surveys could actually be observing two different populations. Given RV surveys' detec-

tion bias towards more massive planets and transit surveys' bias towards larger planets, HARPS could be detecting dense super-Earths while *Kepler* could be detecting volatile-rich sub-Neptunes.

Ideally, assumptions about the bulk densities present in a population of planets would be informed by observations of planets that are detected by both methods. Unfortunately, contemporary observations of sub-Neptune-mass planets could not yet provide robust constraints on this issue: Kepler-11 b - f (Lissauer et al., 2011a), whose mass measurements have significant errors ($\pm$ 30 - 100%), were the only confirmed transiting planets (as of the submission of the paper corresponding to this chapter, Wolfgang & Laughlin 2012) that fell securely in the mass and period ranges considered here ($1 \leq M \leq 17$ $M_\oplus$ and $2 \leq P \leq 50$ days). With the underlying planet population's bulk density distribution so poorly constrained, we felt it necessary to investigate how different density distributions, via different population-wide mass-to-radius relationships (M-Rs), affect the total number of planets that can be detected by either HARPS or *Kepler*, with an eye on how this informs the interpretation of their occurrence rates.

We first investigated the power-law M-R observed in the Solar System, $R/R_\oplus = (M/M_\oplus)^{0.48}$, the same one used in Lissauer et al. (2011b) to test the long-term stability of multiple planetary systems. This single-valued M-R automatically incorporates the mass-density trend between the smaller terrestrial planets and the larger ice giants in our Solar System. When applied to the HARPS super-Earth occurrence rate, this M-R produces a locus of parameter space where a planet population could produce the total number of *Kepler* planet candidates observed in its first four months of data (see Figure

2.3). However, the overall fit of this M-R is poor (§2.4.1), and so we turned to an M-R that does not assume our Solar System provides sufficient insight into the possible densities of planets that have no Solar System analogs. In particular, we investigate the effect of allowing a range of densities to occur at each planet mass through our multi-valued M-R (§2.3.2).

Hints of a multi-valued M-R have been present since the first two dually-detected "Super-Earth"s were measured to have similar masses but significantly different densities (CoRoT-7 b: 5.6 g/cm^3 at 4.8 ± 0.8 M$_\oplus$, Queloz et al., 2009, Léger et al., 2009; GJ 1214 b: 1.9 g/cm^3 at 6.55 ± 0.98 M$_\oplus$, Charbonneau et al., 2009). These hints have continued to emerge with more recent detections: most of the Kepler-11 planets have low bulk densities (0.5 - 3.1 g/cm^3; Lissauer et al., 2011a), while Kepler-10 b and 55 Cnc e yield densities of 9 g/cm^3 (Batalha et al., 2011) and 5 - 6 g/cm^3 (Winn et al., 2011; Demory et al., 2011), respectively. A popular explanation for this compositional bimodality is that the high-density planets, which so far are all observed on extremely close-in orbits ($P < 2$ days), constitute the special case of low-mass gas planets that have had their atmospheres completely stripped, leaving only their solid cores behind (Schaefer & Fegley, 2009; Jackson et al., 2010; Batalha et al., 2011). Instead, we propose that these high-density planets might constitute a more general short-period — and thus more easily detectable — case of an entirely different class of exoplanets: true super-Earths that formed with a primarily refractory composition. This new interpretation has significant implications for planet formation (i.e. Hansen & Murray, 2012), suggesting that there may be multiple modes of formation for planets in this mass range (Léger

et al., 2011).

The multi-valued M-R we present here (§2.3.2) adopts two compositions: rocky planets that follow the same $R/R_\oplus = (M/M_\oplus)^{0.33}$ relationship as the Solar System's inner planets, and gaseous planets that follow the M-R curves presented in Rogers et al. (2011), while a prescription for atmospheric mass loss introduces a third intermediate composition. An admixture of these compositions over the entire 1 $M_\oplus \leq M \leq$ 17 $M_\oplus$ mass range is able to account for the density variation currently observed among low-mass planets. We emphasize that the order-of-magnitude mass loss prescription we appeal to here does not attempt to model the details of atmospheric escape; we use it only as a way to account for the evolution of a gaseous planet's radius in the low-mass, large-radius regime. Interestingly, the presence of intermediate-density planets in a period-radius parameter space unoccupied by rocky or gaseous planets suggests that an intermediate-density planet population, however its constituent planets were formed, is another key component of the transit-RV comparison.

For our multi-valued M-R, we have placed particular emphasis on parameterizing the relative contributions from the rocky and gaseous compositions in as physically intuitive a way as possible, while taking care to minimize the number of free parameters. As a result, we adopt a parameterization that flows naturally from the coexistence of rocky super-Earths and gaseous sub-Neptunes and involves only two additional degrees of freedom: (1) the fraction of all 1 $M_\oplus$ planets in the simulated planet population that have a rocky composition, $f_{rocky}(1)$, and (2) the fraction of all 17 $M_\oplus$ planets that have a gaseous composition, $1 - f_{rocky}(17)$, with f_{rocky} varying linearly between the bounding

masses.

As seen in Figure 2.5, this multi-valued M-R also produces agreement between the total number of *Kepler* planet candidates and the total number of detectable planets simulated from the HARPS occurrence rate, albeit for slightly different mass and period distributions than the single-valued M-R (§2.4.2). Thus, constraints from RV surveys on the super-Earth mass distribution can be combined with this investigation to rule out M-Rs that do not fully describe the observed planet population in a global sense.

When the details of the simulated period-radius distributions are compared against *Kepler*'s, however, we see that the multi-valued M-R still does not sufficiently describe *Kepler*'s planet candidates (§2.4.2). This is not unexpected considering how many variables affect the final compositions and orbital configurations of planets around completely different stars in different environments. Given the simplifying assumptions we applied to the input period and mass distributions (§2.3.1) and to our prescription for atmospheric mass loss (§2.3.2), we would have been surprised if our simple 4-parameter model was able to fully explain the diversity of low-mass planets *Kepler* has discovered. It is likely that other population-wide M-Rs which do take these variables into account could improve upon our transit-RV fit; however, it remains unclear whether these M-Rs would be parameterizable in a reasonable number of degrees of freedom. In this investigation, we have chosen simplicity over absolute best fits in an effort to study M-Rs that are readily accessible and physically intuitive; correspondingly, we show only that, for certain parameter sets, both M-Rs can explain the apparent discrepancy between the total number of planets that *Kepler* has observed and the HARPS overall super-

Earth occurrence rate, not the details of the *Kepler* planet candidate period-radius distribution.

2.5.1 Caveats

Before one concludes that certain parameter sets illustrated in Figures 2.3 and 2.5 can resolve the occurrence rate discrepancy between *Kepler* and HARPS, a number of other factors must be considered. First, by using the *Kepler* planet candidates directly, we have implicitly assumed that the *Kepler* Science Processing Pipeline is very close to 100% complete. The completeness of the transit search algorithm is a separate issue from the detection completeness that we account for in §2.3.4: a 100% pipeline completeness assumes that every planet which *Kepler* was able to detect in its first four months of data was actually detected. If *Kepler*'s list of planet candidates is not complete in this sense, then the number we compare N_{detect} to would rise, pushing the parameter sets which produce agreement to those with an orange or yellow color.

Also of concern is our use of the true planet mass, M, rather than the observationally determined $M\sin(i)$, the quantity that RV surveys like HARPS actually measure. Our choice arose from acknowledging that the inclination distribution of observed radial velocity planets is poorly understood and that spherical isotropy cannot be assumed due to the detection biases inherent in the radial velocity technique. However, this assumption produces a systematic effect on N_{detect} that we must consider. Because assuming M instead of $M\sin(i)$ underestimates the observed planets' true masses, we include planets that would otherwise fall outside of our considered mass range, resulting in a true N_{detect} that is less than the values we report here. Although we cannot estimate

the magnitude of this effect because we do not know the distribution of i, it qualitatively opposes the effect that the false positive rate has on the transit-RV comparison. Thus, the values of $N_{detect} \sim 630$ produced by both M-Rs appear to remain consistent with the total number of *Kepler*'s planet candidates.

Third, we must carefully examine the different definitions of "occurrence rate" implemented in transit and RV surveys. The HARPS overall occurrence rate, which makes a statement about the fraction of stars with planets, treats the presence of planets around stars as a binary state: either the star hosts no planets, or it hosts one or more planets. Because the HARPS occurrence rate offers no information about the appropriate multiple-planet assumptions to make, we choose to restrict each host star to only one planet (§2.3.1); we account for this by only considering the first planet candidate in our radius and period range to be listed by the *Kepler* pipeline in each multiple-planet system. On the other hand, the *Kepler* occurrence rates computed by Howard et al. (2012) and Youdin (2011) include the possibility of multiple-planet systems and give the number of planets per star (NPPS), rather than the fraction of stars with planets (FSWP). With information about the distribution of multiple-planet systems such as that offered by Latham et al. (2011) and Tremaine & Dong (2012), an NPPS occurrence rate can be directly compared to a FSWP occurrence rate. For our purposes we simply note that the occurrence rates which Howard et al. (2012) and Youdin (2011) compute (0.13 and 0.19 planets per star, respectively, for $2 \leq R < 4$ $R_\oplus$ and $P < 50$ days) would become even lower when transformed to a FSWP occurrence rate, given the presence of multiple-planet systems. As seen by the $\sim 20\%$ overall reduction of included

planet candidates produced by our single-planet assumption (§2.3.5), this only worsens the apparent discrepancy between the two surveys' occurrence rates. Youdin (2011) does point out, however, that if planets down to 0.5 $R_\oplus$ are included, then this number-of-planets-per-star occurrence rate may be as high as 1.36. Thus, for the full $1 \leq R < 4$ $R_\oplus$ range we consider in this chapter, the apparent occurrence rate discrepancy may also be explained at least in part by the slight differences in the considered radius range.

Fourth, significant errors in the *Kepler* target stars' radii will affect the total number of true *Kepler* planets in the radius range we consider here. Assuming normally distributed errors, the $\sim$ 10% uncertainty in the target stars' radii — and thus the planet candidates' radii — produced by the uncertainty in the Kepler Input Catalog's estimates of T_{eff} and $\log(g)$ (Brown et al., 2011) is not enough to appreciably change the total number of planet candidates that we compare our simulations to. However, it is probable that unaccounted-for systematic errors are at play in R_{pl}. Indeed, the KIC radii are known to be severely biased in at least one instance: the presence of unidentified subgiant stars in the target stars list can underestimate the stellar radii by as much as a factor of 2 (Brown et al., 2011). We have attempted to minimize the effect of such a severe systematic error by limiting the *Kepler* target stars we consider to only those with $\log(g) > 4.0$ (§2.3.3), but this does not guarantee that our sample of potential host stars are completely free of systematic biases that could change the total number of planet candidates we use for our transit-RV comparison.

A final but significant source of concern is the difference between each survey's target star selection criteria. We address the biases produced from *Kepler*'s selection

criteria by drawing from the Q2 targets stars, and we account for its detection incompleteness by including the Q2 3-hour CDPP measurements; these considerations stem from how we frame the transit-RV comparison, as we ask how many short-period, low-mass planets *Kepler* would have detected in its first four months of data if the HARPS occurrence rate is true. To make a thorough comparison, however, one also needs to consider how these biases differ from the RV selection criteria that factor into HARPS' overall occurrence rate. Both HARPS and *Kepler* preferentially choose G and K dwarfs with high signal-to-noise ratios (Mayor et al., 2009; Udry et al., 2000; Batalha et al., 2010b), but HARPS also targets slowly rotating, magnetically quiet stars and includes no known spectroscopic binaries. Thus, the differences between the two survey's selection criteria lie in the presence of binary stars in the *Kepler* sample and in the distinction between RV stellar jitter and photometric noise.

According to Batalha et al. (2010b), *Kepler* searches for planets around all of the known eclipsing binaries (> 600) in its field of view. While these eclipsing binaries are not numerous enough by themselves to appreciably affect our statistics, the unidentified spectroscopic binaries in *Kepler*'s field of view potentially are, if one reasonably allows for the possibility that the planet occurrence rate can differ between single stars and binary systems. To get a sense for the magnitude of this effect, we refer to Duquennoy & Mayor (1991), who estimate that as many as two thirds of all G dwarfs have a stellar companion. The lognormal period distribution they find for spectroscopic G-dwarf binaries indicates that roughly 8% of all G dwarfs exist in binaries separated by < 0.5 AU and $\sim 20\%$ in binaries separated by $\lesssim 10$ AU; considering that *Kepler*'s

false-positive vetting process enables binaries at separations of $< 1''$ (Batalha et al., 2011) to be identified, the relative fraction of tight binaries in the *Kepler* target star list could be even higher. Separations of < 0.5 AU and < 10 AU are especially of interest for the survival and formation of planets in binary systems, as the orbits of the planets considered in this chapter would not be stable in equal-mass binary systems separated by < 0.5 AU, and protoplanetary disks around the primaries of $\lesssim 10$ AU binary systems would be truncated before the distance at which an ice line could form, assuming an equal-mass binary of solar-type stars and thus an ice line at a distance of approximately 5 AU. Interestingly, a difference in the planet occurrence rate for binaries with < 10 AU separations versus those with > 10 AU separations could provide a way to discriminate between the compositions of these close-in planets, if the terrestrial planets formed in-situ and the gaseous planets migrated in from wider orbits.

The HARPS requirement that its target stars have low levels of RV stellar jitter is another potentially significant difference between the two surveys' target selection criteria. It is certainly the case that *Kepler* has preferentially chosen target stars that exhibit low photometric noise (Batalha et al., 2010b), but this noise is primarily correlated with the apparent magnitude of the star (i.e. Figure 2.1) and does not necessarily reflect the degree of magnetic activity that heavily factors into the HARPS $\log(R'_{HK}) < -4.8$ target selection. If we temporarily ignore this, however, and assume that photometric noise is strongly correlated with stellar jitter, we can assess the effect of this selection criterion on our results. We find that limiting our potential host stars to the $\sim$ 35,000 *Kepler* targets with $\mathrm{CDPP}_3 \leq 150$ ppm worsens the discrepancy be-

tween the *Kepler* and HARPS occurrence rates: for $\alpha = -1.0$, $\beta = 0.0$, $f_{rocky}(1) = 0.9$, $f_{rocky}(17) = 0.1$, and a 40% overall occurrence rate, we find that *Kepler* would have been able to detect 291 ± 19 planets in its first four months of data ($N_{realizations} = 100$), while *Kepler* has actually found 217 planet candidates around stars with $CDPP_3 \leq 150$ ppm. A 30% HARPS occurrence rate is needed to bring these numbers into agreement, making the HARPS-*Kepler* consistency marginal at best, although a high spectroscopic binary fraction in the *Kepler* sample could counteract this effect. In any case, systematically accounting for the selection of quiet stars requires the forthcoming results of stellar photometric variation studies (i.e. Basri et al., 2011) to draw conclusions about the *Kepler* target stars' magnetic activity, given the absence of spectra for a majority of these targets.

In short, we acknowledge that the differences in the two surveys' target star selection criteria could explain some of the apparent discrepancy between their occurrence rates. Our intent here is simply to point out plausible, testable explanations for an overall transit-RV occurrence rate discrepancy that does not depend on the selection criteria to produce similar numbers of observable planets.

2.6 Conclusion

In summary, we investigate the effect that two different mass-to-radius relationships have on the *Kepler*-HARPS comparison, in terms of the total number of planets detectable by *Kepler* in its first four months of data. Both the single-valued M-R and the multi-valued M-R can bring the apparent discrepancy between the occur-

rence rates into alignment, but for different mass and period distributions; this enables future observations to rule out M-Rs that do not fully describe the observed planet population in a global sense. Furthermore, we present for the first time a multi-valued M-R, the existence of which is hinted at by direct observations, that follows naturally from simultaneously extending Neptune-like planets to lower masses and Earth-like planets to higher masses. By illustrating how a number of parameter sets at realistic values of $f_{rocky}(1)$, and $f_{rocky}(17)$ can produce N_{detect} consistent with the total number of analogous *Kepler* planet candidates, we show that this M-R can provide a physically intuitive explanation for the apparent discrepancy between RV and transit surveys' planet occurrence rates, wherein HARPS may be detecting a large population of dense low-mass planets and *Kepler* may be detecting a large population of gaseous sub-Neptunes.

Chapter 3

Probabilistic Mass-Radius Relation for Sub-Neptune-Sized Planets

3.1 Introduction

The emergence of the sub-Neptune population, which has no Solar System analogs, poses fundamental questions about the typical compositional constituents of planets within a few times Earth's size. As bulk densities offer some insight into this problem, these planets' individual mass and radius measurements provide observational constraints for theoretical composition studies. Recently these studies have shifted to considering the available planets as a statistical ensemble (e.g. Rogers 2015; Wolfgang & Lopez 2015 sans mass constraints), which motivates detailed analyses of the observed

mass-radius distribution.

The joint mass-radius distribution, which is often couched in terms of the mass-radius "relationship" (M-R relation), is also highly relevant for dynamical and formation studies of the *Kepler* planet candidates (PCs). Mass measurements for individual PCs are often unavailable, as the majority orbit stars too faint for Doppler follow-up (Batalha et al., 2010b) and only $\sim 6\%$ exhibit transit timing variations (TTVs) at high signal-to-noise ratios (Mazeh et al., 2013). Therefore, a statistical "conversion" is necessary to map observed radii to the masses these studies need.

To date, several M-R relations have been posed in the exoplanet literature. To solve the practical issue described above, Lissauer et al. (2011b) fit a power law to Earth and Saturn and found $M = R^{2.06}$, where M and R are in Earth units. Wu & Lithwick (2013) derived masses using the amplitudes of sinusoidal TTVs for 22 planet pairs, and found $M = 3R$. More recently, Weiss & Marcy (2014), hitherto WM14, fit a power law to masses and radii available in the literature, which was dominated by the 42 planets chosen by the *Kepler* team to be followed up with radial velocity measurements (Marcy et al., 2014); they found $M = 2.69R^{0.93}$ for planets with $1.5 < R < 4$ $R_\oplus$.

All of these results were produced via basic least squares regression, which is commonly used in astronomy to fit lines through points. However, this classic technique does not properly account for several issues that are relevant to the small-planet M-R relation: measurement uncertainty in the independent variable (i.e. planet radii), non-detections and upper limits, and intrinsic, astrophysical scatter in the dependent variable (i.e. planet masses). Thankfully, there are solutions to these problems in both

the Bayesian and frequentist statistics literature (see §1 of Kelly (2007) for a concise overview). We present an example of one of these techniques which can be executed using existing numerical algorithms and code (§5.3.3), which is effectively a simplified implementation of the Kelly (2007) linear regression scheme.

Of particular interest is the intrinsic scatter that has not been previously characterized. Theoretical work on planet compositions suggest this scatter should exist: thermally evolved rock-hydrogen sub-Neptune internal structure models yield radii mostly independent of mass (Lopez & Fortney, 2014), which produces significant mass-radius scatter when a distribution of gaseous mass fractions is present in the population (Wolfgang & Lopez, 2015). Furthermore, the mere presence of otherwise layered exoplanets produces a range of radii at a given mass due only to differences in the layers' compositions (e.g. Seager et al., 2007; Fortney et al., 2007; Rogers et al., 2011). This motivates us to move beyond deterministic, one-to-one mappings, which are in a sense "mean" relationships and which cause studies that use them to be accurate on average only. This average accuracy is insufficient and inappropriate if one's aim is to argue for a particular physical process based on full distributions of parameters (versus qualitative comparison to observations), or if the purpose is to rule out parts of parameter space, which requires knowledge of the full mass-radius distribution. Indeed, recent formation studies have already begun to fit probabilistic density distributions to observed masses and radii (Chatterjee & Ford, 2015).

In this chapter we show how a probabilistic M-R relation can be constructed (§3.2) and constrained (§5.3.3) using any subset of planetary masses and radii (§3.3).

We also highlight the observational evidence for this expected intrinsic scatter and quantify it in a statistically robust way that includes uncertainties on the M-R relation parameters (§3.5). We discuss the correct usage and some major implications of these findings in §3.6.

3.2 Modeling the M-R Relation

Power laws are often used to parameterize the M-R relation because they are conceptually and computationally simple and can be easily fit to data using the familiar tool of linear regression. We continue with this choice to facilitate more direct comparisons with previous work and to illustrate how a hierarchical framework enables straightforward extensions to entire families of M-R relations. In addition, we cast this in terms of $M(R)$ instead of $R(M)$ to address the practical problem of estimating masses from *Kepler* radii.

In particular, we consider three power law-based M-R relations (Eqns 3.1-3.3). The first is the form used by most prior studies (see §5.1):

$$\frac{M}{M_\oplus} = C\Big(\frac{R}{R_\oplus}\Big)^\gamma \tag{3.1}$$

where M is the mass of the planet, R is the planetary radius, and C and γ are the parameters to be fit to the data. This relation is deterministic in the sense that only one mass is allowed for a given radius.

If instead we want to allow for a range — that is, if we want to incorporate the expected intrinsic scatter — then we need to create an M-R relation which specifies

how those masses should be distributed at a given input radius. Again, taking the most simple, familiar, and analytically tractable approach, we choose a Gaussian distribution, where the mean population mass μ is given by the above power-law relation and where the standard deviation σ_M (units of $M_\oplus$) parameterizes the intrinsic scatter in planet masses:

$$\frac{M}{M_\oplus} \sim \text{Normal}\Big(\mu = C\Big(\frac{R}{R_\oplus}\Big)^\gamma, \sigma = \sigma_M\Big) \tag{3.2}$$

Note that $\sim$ means "drawn from the distribution", thereby marking the difference between a deterministic and a probabilistic M-R relation. Figure 3.1 is the graphical model corresponding to Eqn 3.2, and includes Gaussian error bars on the measured masses and radii (see §5.3.3 for all details of the model).

Generalizing further, the width of the intrinsic scatter may change as planets increase in size, so we consider a probabilistic M-R relation that allows the standard deviation itself to vary as a function of radius via the slope β (units of $M_\oplus^2 R_\oplus^{-1}$):

$$\frac{M}{M_\oplus} \sim \text{Normal}\Big(\mu = C\Big(\frac{R}{R_\oplus}\Big)^\gamma, \sigma = \sqrt{\sigma_{M1}^2 + \beta\tilde{R}}\Big) \tag{3.3}$$

where $\tilde{R} = R/R_\oplus - 1$ and σ_{M1} is now the standard deviation in planet masses at 1 $R_\oplus$ ($\tilde{R} = 0$).

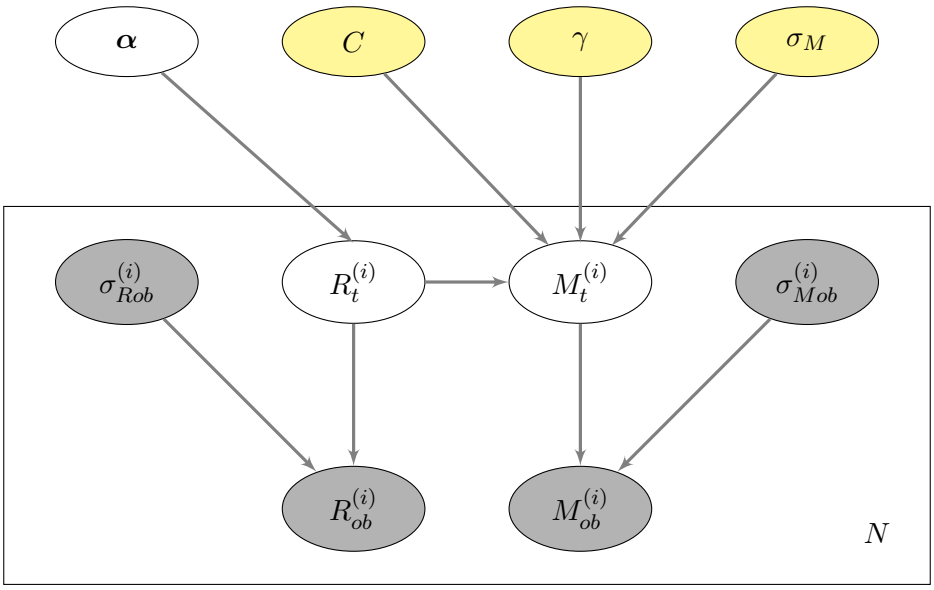

Figure 3.1: Graphical model used to find the best-fit parameters for the probabilistic mass-radius relationship in Eqn 3.2. These parameters of interest are yellow while the observed data are gray (see §3.3) and unobserved parameters are white; definitions are below. Explicitly including the unobserved true masses and radii in the model allow us to easily incorporate physical constraints (such as requiring $M_t^{(i)} > 0$) while preserving all of the information contained in the data (which may allow $M_{ob}^{(i)}$ to span 0). The full model with details of the utilized probability distributions is displayed in Equation 5.9.

$\boldsymbol{\alpha}$ = hyperparameters on radius distribution (see §5.3.3)
C = constant in mean M-R relation
γ = power law index of mean M-R relation
σ_M = intrinsic dispersion in planet masses at a given radius
$R_t^{(i)}$ = true radius of the ith planet
$R_{ob}^{(i)}$ = observed radius of the ith planet
$\sigma_{Rob}^{(i)}$ = measurement uncertainty in $R_{ob}^{(i)}$
$M_t^{(i)}$ = true mass of the ith planet
$M_{ob}^{(i)}$ = observed mass of the ith planet
$\sigma_{Mob}^{(i)}$ = measurement uncertainty in $M_{ob}^{(i)}$

3.3 Data

With the statistical M-R relations defined, we turn to the problem of identifying which dataset to use. Optimally we would use a subset of mass and radius measurements that is uniform and complete, as any systematic biases present in the sample will manifest as biased M-R parameter values. Unfortunately, the available masses and radii are far from this ideal, with mass measurements made with two fundamentally different methods by many different pipelines and chosen for follow-up by a complex, poorly documented selection function. There is significant work to be done to understand how these systematics affect the M-R relation, but it is outside the scope of this chapter, as our main purpose is to show how a probabilistic M-R relation can be derived from whichever dataset one wishes to use. Therefore, we choose a baseline dataset consisting of radial velocity-measured masses, which somewhat reduces the heterogeneity of the sample while preserving a fairly large number of data points.

Table 3.1 shows all of the masses and radii that we consider, with our baseline dataset denoted with a label of 0; the list was constructed by starting with the WM14 dataset and identifying new planets and updates in the NASA Exoplanet Archive (last accessed 1/30/2015). We manually double-checked each planet to verify that the reported measurements were correct and most up-to-date, paying particular attention to which methods and stellar parameters were used (data denoted by a label of 1 were present in and haven't changed since WM14). The TTV dataset (label of 2) contains only the sub-Neptune-sized planets with photodynamical models fit to their transit timing variations, as these masses are the best constrained and therefore provide the most

Table 3.1: Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
55 Cnc e	0.737	8.09	0.26	2.17	0.098	McArthur (2004)	Nelson(2014), Gillon(2012)	0
CoRoT-7 b	0.854	4.73	0.95	1.58	0.064	Queloz(2009); Leger(2009)	Barros(2014)	0
GJ 1214 b	1.580	6.45	0.91	2.65	0.09	Charbonneau (2009)	Carter(2011)	0,1
GJ 3470 b	3.337	13.73	1.61	3.88	0.32	Bonfils(2012)	Biddle(2014)	0
HD 97658 b	9.491	7.87	0.73	2.34	0.16	Howard(2011)	Dragomir (2013)	0,1
HIP 116454 b	9.12	11.82	1.33	2.53	0.18	Vanderburg (2015)	Vanderburg (2015)	0 0
Kepler-10 b	0.837	3.33	0.49	1.47	0.02	Batalha(2011)	Dumusque (2014)	0
Kepler-10 c	45.294	17.2	1.9	2.35	0.06	Batalha(2011)	Dumusque (2014)	0
Kepler-19 b	9.287	-99	20.3	2.21	0.048	Borucki(2011)	Ballard(2011)	0,4
Kepler-20 b	3.696	8.7	2.2	1.91	0.16	Borucki(2011)	Gautier(2012)	0
Kepler-20 c	10.854	16.1	3.5	3.07	0.25	Borucki(2011)	Gautier(2012)	0
Kepler-20 d	77.612	-99	20.1	2.75	0.23	Borucki(2011)	Gautier(2012)	0,4
Kepler-20 e	6.098	-99	3.08	0.868	0.08	Borucki(2011)	Fressin(2012)	0,4

information for the sub-Neptune M-R relation; neither circumbinary planets nor unconfirmed planets were included, again to try to keep a somewhat more homogeneous dataset. Finally, to enable easier comparison with previous work, we continued the error treatment of WM14: if asymmetric upper and lower uncertainties were reported, we used their average as a symmetric 1σ error bar[1]. 2σ upper limits were included if they were < 80 $M_\oplus$ for $R < 4$ $R_\oplus$ and < 300 $M_\oplus$ for $4 < R < 8$ $R_\oplus$.

[1]Future work using HBM can improve on this error treatment by using the full posteriors of the mass and radius measurements, if these posteriors are made available in the literature.

Table 3.1 (cont'd): Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
Kepler-20 f	19.58	-99	14.3	1.03	0.11	Borucki(2011)	Fressin(2012)	0,4
Kepler-21 b	2.786	-99	10.4	1.635	0.04	Borucki(2011)	Howell(2012)	0,4
Kepler-25 b	6.239	9.60	4.20	2.71	0.05	Borucki(2011)	Marcy(2014)	0,1
Kepler-37 b	13.367	2.78	3.70	0.32	0.02	Borucki(2011)	Marcy(2014)	0,1
Kepler-37 c	21.302	3.35	4.00	0.75	0.03	Borucki(2011)	Marcy(2014)	0,1
Kepler-37 d	39.792	1.87	9.08	1.94	0.06	Borucki(2011)	Marcy(2014)	0,1
Kepler-48 b	4.778	3.94	2.10	1.88	0.10	Borucki(2011)	Marcy(2014)	0,1
Kepler-48 c	9.674	14.61	2.30	2.71	0.14	Borucki(2011)	Marcy(2014)	0,1
Kepler-48 d	42.896	7.93	4.60	2.04	0.11	Borucki(2011)	Marcy(2014)	0,1
Kepler-62 b	5.715	-99	9	1.31	0.04	Borucki(2011)	Borucki(2013)	0,4
Kepler-62 c	12.44	-99	4	0.54	0.03	Borucki(2013)	Borucki(2013)	0,4
Kepler-62 d	18.164	-99	14	1.95	0.07	Borucki(2011)	Borucki(2013)	0,4
Kepler-62 e	122.39	-99	36	1.61	0.05	Borucki(2011)	Borucki(2013)	0,4
Kepler-62 f	267.29	-99	35	1.41	0.07	Borucki(2013)	Borucki(2013)	0,6
Kepler-68 b	5.399	5.97	1.70	2.33	0.02	Borucki(2011)	Marcy(2014)	0,3
Kepler-68 c	9.605	2.18	3.50	1.00	0.02	Batalha(2013)	Marcy(2014)	0,3
Kepler-78 b	0.354	1.69	0.41	1.20	0.09	Sanchis-Ojeda(2013a)	Howard(2013)	0,1
Kepler-89 b	3.743	10.50	4.60	1.71	0.16	Borucki(2011)	Weiss(2013)	0,1
Kepler-93 b	4.727	4.02	0.68	1.48	0.019	Borucki(2011)	Dressing(2015)	0

Table 3.1 (cont'd): Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
Kepler-94 b	2.508	10.84	1.40	3.51	0.15	Borucki(2011)	Marcy(2014)	0,1
Kepler-95 b	11.523	13.00	2.90	3.42	0.09	Borucki(2011)	Marcy(2014)	0,1
Kepler-96 b	16.238	8.46	3.40	2.67	0.22	Borucki(2011)	Marcy(2014)	0,1
Kepler-97 b	2.587	3.51	1.90	1.48	0.13	Borucki(2011)	Marcy(2014)	0,1
Kepler-98 b	1.542	3.55	1.60	1.99	0.22	Borucki(2011)	Marcy(2014)	0,1
Kepler-99 b	4.604	6.15	1.30	1.48	0.08	Borucki(2011)	Marcy(2014)	0,1
Kepler-100 b	6.887	7.34	3.20	1.32	0.04	Borucki(2011)	Marcy(2014)	0,1
Kepler-100 c	12.816	0.85	4.00	2.20	0.05	Borucki(2011)	Marcy(2014)	0,1
Kepler-100 d	35.333	-4.36	4.10	1.61	0.05	Borucki(2011)	Marcy(2014)	0,1
Kepler-101 c	6.03	-99	9	1.25	0.18	Borucki(2011)	Bonomo(2014)	0,5
Kepler-102 d	10.312	3.80	1.80	1.18	0.04	Borucki(2011)	Marcy(2014)	0,1
Kepler-102 e	16.146	8.93	2.00	2.22	0.07	Borucki(2011)	Marcy(2014)	0,1
Kepler-102 f	27.454	0.62	3.30	0.88	0.03	Borucki(2011)	Marcy(2014)	0,1
Kepler-102 b	5.287	0.41	1.60	0.47	0.02	Borucki(2011)	Marcy(2014)	0,1
Kepler-102 c	7.071	-1.58	2.00	0.58	0.02	Borucki(2011)	Marcy(2014)	0,1
Kepler-103 b	15.965	14.11	4.70	3.37	0.09	Borucki(2011)	Marcy(2014)	0,1
Kepler-106 b	6.165	0.15	2.80	0.82	0.11	Borucki(2011)	Marcy(2014)	0,1
Kepler-106 c	13.571	10.44	3.20	2.50	0.32	Borucki(2011)	Marcy(2014)	0,1
Kepler-106 d	23.980	-6.39	7.00	0.95	0.13	Batalha(2013)	Marcy(2014)	0,1
Kepler-106 e	43.844	11.17	5.80	2.56	0.33	Borucki(2011)	Marcy(2014)	0,1

Table 3.1 (cont'd): Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
Kepler-109 b	6.482	1.30	5.40	2.37	0.07	Borucki(2011)	Marcy(2014)	0,1
Kepler-109 c	21.223	2.22	7.80	2.52	0.07	Borucki(2011)	Marcy(2014)	0,1
Kepler-113 b	4.754	7.10	3.30	1.82	0.05	Borucki(2011)	Marcy(2014)	0,1
Kepler-113 c	8.925	-4.60	6.20	2.19	0.06	Borucki(2011)	Marcy(2014)	0,1
Kepler-131 b	16.092	16.13	3.50	2.41	0.20	Borucki(2011)	Marcy(2014)	0,1
Kepler-131 c	25.517	8.25	5.90	0.84	0.07	Batalha(2013)	Marcy(2014)	0,1
Kepler-406 b	2.426	4.71	1.70	1.43	0.03	Borucki(2011)	Weiss(2014)	0,1
Kepler-406 c	4.623	1.53	2.30	0.85	0.03	Batalha(2013)	Weiss(2014)	0,1
Kepler-407 b	0.669	0.06	1.20	1.07	0.02	Borucki(2011)	Marcy(2014)	0,1
Kepler-409 b	68.958	2.69	6.20	1.19	0.03	Batalha(2013)	Marcy(2014)	0,1
Kepler-4 b	3.213	24.47	3.81	4.00	0.21	Borucki(2010)	Borucki(2010)	
GJ 436 b	2.64	25.4	2.1	4.10	0.16	Butler(2004)	Lanotte(2014)	
Kepler-89 c	10.42	15.6	10.6	4.32	0.41	Batalha(2013)	Weiss(2013)	
HAT-P-11 b	4.888	25.74	2.86	4.73	0.157	Bakos(2010)	Bakos(2010)	
CoRoT-22 b	9.756	-99	35	4.88	0.28	Moutou(2014)	Moutou(2014)	4
Kepler-103 c	179.61	36.1	25.2	5.14	0.14	Borucki(2011)	Marcy(2014)	
Kepler-101 b	3.488	51.1	4.9	5.77	0.82	Borucki(2011)	Bonomo(2014)	
Kepler-63 b	9.43	-99	95	6.1	0.2	Borucki(2011)	Sanchis-Ojeda(2013b)	6
HAT-P-26 b	4.235	18.75	2.23	6.33	0.58	Hartman(2011)	Hartman(2011)	

Table 3.1 (cont'd): Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
CoRoT-8 b	6.212	69.92	9.53	6.39	0.22	Borde(2010)	Borde(2010)	
Kepler-89 e	54.32	35	23	6.56	0.62	Batalha(2013)	Weiss(2013)	
Kepler-11 b	10.304	1.90	1.2	1.80	0.04	Lissauer(2011)	Lissauer(2013)	1,2
Kepler-11 c	13.024	2.90	2.3	2.87	0.06	Lissauer(2011)	Lissauer(2013)	1,2
Kepler-11 d	22.684	7.30	1.2	3.12	0.07	Lissauer(2011)	Lissauer(2013)	1,2
Kepler-11 f	46.689	2.00	0.9	2.49	0.06	Lissauer(2011)	Lissauer(2013)	1,2
Kepler-11 g	118.38	-99	25	3.33	0.07	Lissauer(2011)	Lissauer(2013)	2,4
Kepler-18 b	3.505	6.9	3.4	2.00	0.100	Borucki(2011)	Cochran(2011)	1,2
Kepler-30 b	29.334	11.3	1.4	3.90	0.200	Borucki(2011)	Sanchis-Ojeda(2012)	1,2
Kepler-36 b	13.840	4.45	0.30	1.486	0.035	Carter(2012)	Carter(2012)	1,2
Kepler-36 c	16.239	8.08	0.53	3.679	0.054	Borucki(2011)	Carter(2012)	1,2
Kepler-79 b	13.485	10.9	6.7	3.47	0.07	Borucki(2011)	Jontof-Hutter(2014)	1,2
Kepler-79 c	27.403	5.9	2.1	3.72	0.08	Borucki(2011)	Jontof-Hutter(2014)	1,2
Kepler-79 e	81.066	4.1	1.2	3.49	0.14	Batalha(2013)	Jontof-Hutter(2014)	1,2
Kepler-88 b	10.954	8.7	2.5	3.78	0.38	Borucki(2011)	Nesvorny(2013)	2

Table 3.1 (cont'd): Masses and Radii of Small Planets

Planet Name	Period (days)	M_{obs} ($M_\oplus$)	σ_{Mobs} ($M_\oplus$)	R_{obs} ($R_\oplus$)	σ_{Robs} ($R_\oplus$)	First Reference	Mass, Radius Reference	Note
Kepler-138 c	13.782	3.83	1.39	1.610	0.160	Borucki(2011)	Kipping(2014)	2
Kepler-138 d	23.089	1.01	0.38	1.610	0.160	Borucki(2011)	Kipping(2014)	2
Kepler-289 b	34.545	7.3	6.8	2.15	0.1	Borucki(2011)	Schmitt(2014)	2
Kepler-289 d	66.063	4.0	0.9	2.68	0.17	Borucki(2011)	Schmitt(2014)	2

Note. —
0. Included in baseline dataset.
1. Mass, radius values and their error bars are unchanged (within rounding error) from WM14.
2. Mass measured via a full photodynamical fit to TTVs.
3. The Kepler-68 planets were repeated twice in the WM14 dataset, so we use the Marcy et al. (2014) values.
4. The σ_{Mobs} column contains the 2σ upper limit as reported in the second reference.
5. Only a 1σ upper limit of 3.78 was given, and no posteriors were shown; in this analysis, we set the 2σ upper limit at 9 $M_\oplus$ to include 1.8 m/s uncertainty quoted in RV semi-amplitude for the larger Kepler-101 b.
6. The 2σ upper limit is interpolated from given 1σ and 3σ upper limits.

3.4 Fitting the M-R Relations

We use hierarchical Bayesian modeling (HBM) to fit the M-R relations in §3.2 to the data described in §3.3. This statistical method is described in detail in Wolfgang & Lopez (2015) in the context of exoplanet compositions; further pedagogical discussion and examples of HBM in the astronomical literature is provided by Loredo (2013). A very similar approach to this HBM-enabled linear regression was detailed in Kelly (2007); we refer the reader to that paper for an in-depth discussion of the general advantages and improvements of this approach over the commonly used χ^2 analysis for linear regression.

For the problem at hand, HBM (or the analogous frequentist methods for multi-level modeling) is necessary for a number of reasons:

- It allows us to directly model and fit the astrophysical dispersion in the population as an explicit parameter.

- It allows us to self-consistently incorporate uncertainties on the independent variable (radii in this case), without the need for elaborate bootstrapping schemes.

- Most sub-Neptune mass uncertainties are large, and some are realistically only upper limits. HBM is able to simultaneously use all likelihood distributions no matter their width or shape, which increases the information content of the resulting M-R relation and decreases the biases that binning or weighting schemes introduce when these likelihoods are asymmetric.

- Relatedly, HBM allows us to introduce the true masses and radii as latent (unobserved) parameters; this enables us to restrict the masses to physically allowed parameter space (such as $M > 0$ or $\rho < \rho_{iron}(M)$) while preserving all of the information in the observations (including the negative mass measurements that are allowed by the data).

- As with all Bayesian methods, HBM produces posterior distributions, allowing us to easily see the uncertainties in the M-R relation parameters. Most of the M-R relations currently reported and used in the literature have no published uncertainties.

The hierarchical model for our default M-R relation (Eqn 3.2) is displayed in Figure 3.1 to clarify the structural relationships between parameters and observables. This structure is also present in the written version below, along with details of the

distributions we used ("N" represents a normal distribution with the listed parameters in order of μ and σ; "U" represents a uniform distribution with the listed numbers bounding the interval; and "|" means "given", i.e. the parameter to the left depends on the parameters to the right):

$$\begin{aligned}
\gamma &\sim \mathrm{N}(1,1) \\
\ln(C) &\sim \mathrm{U}(-3,3) \\
\log(\sigma_M^2) &\sim \mathrm{U}(-4,2) \\
R_t^{(i)} &\sim \mathrm{U}(0.1,10) \\
\mu_M^{(i)}|R_t^{(i)},C,\gamma &= \gamma\ln(R_t^{(i)}) + \ln(C) \\
M_t^{(i)}|R_t^{(i)},C,\gamma,\sigma_M &\sim \mathrm{N}\Big(\mathrm{e}^{\mu_M^{(i)}},\sigma_M\Big) \\
R_{ob}^{(i)}|R_t^{(i)},\sigma_{Rob}^{(i)} &\sim \mathrm{N}(R_t^{(i)},\sigma_{Rob}^{(i)}) \\
M_{ob}^{(i)}|M_t^{(i)},\sigma_{Mob}^{(i)},R_t^{(i)},C,\gamma,\sigma_M &\sim \mathrm{N}(M_t^{(i)},\sigma_{Mob}^{(i)})
\end{aligned} \tag{3.4}$$

For the deterministic M-R relation of Eqn 3.1, Eqn 5.9 remains the same except there is no σ_M parameter, and

$$M_t^{(i)}|R_t^{(i)},C,\gamma = \mathrm{e}^{\mu_M^{(i)}}$$

while for the M-R relation of Eqn 3.3, there was an additional parameter β such that:

$$\beta \sim \mathrm{U}(-10, 10)$$

$$M_t^{(i)} | R_t^{(i)}, C, \gamma, \sigma_{M1}, \beta \sim \mathrm{N}\Big(\mathrm{e}^{\mu_M^{(i)}}, \sqrt{\sigma_{M1}^2 + \beta \tilde{R}_t^{(i)}}\Big)$$

For all M-R relations we consider, we truncated the $M_t^{(i)}$ distribution such that $0 < M_t^{(i)} < \rho_{iron} * (R_t^{(i)})^3$ where $\rho_{iron}(M_t^{(i)})$ was computed using the 0% rock mass fraction analytic fits to the Fortney et al. (2007) rock-iron internal structure models. Additionally, we tested several end member cases for the R_t distribution, and the choice for this prior had a negligible effect on the result, primarily because R_{ob} is fairly well constrained throughout the sample. A wide normal distribution was used in the first line of the model because there was some prior information provided by Wu & Lithwick (2013) and WM14 which indicated that $\gamma \approx 1$ for sub-Neptunes; this distribution is wide enough that a uniform distribution produces very similar results. Note that the normal distributions in the last two lines of the model are the same likelihoods that are assumed when using χ^2 to perform linear regression.

To produce the results shown in §3.5, we evaluate each model with JAGS (Just Another Gibbs Sampler; Plummer 2003), an R code for numerically evaluating hierarchical Bayesian models with MCMC. For each set of posteriors in Figures 3.2 and 3.3, we ran 10 chains consisting of 500,000 iterations each. The first half of each chain is discarded as "burn-in", and the resulting half is thinned by a factor of 250, such that we retain 10,000 posteriors samples of each parameter. JAGS computes the MCMC convergence diagnostic $\hat{R}$ of Gelman & Rubin (1992) at run-time; our models are fully

converged, with all parameters having $\hat{R} \leq 1.002$.

3.5 Results

Table 3.2 shows the results of this modeling; in particular lines $2-6$ show our best-fit parameters for our probabilistic M-R relation (Eqn 3.2) for various datasets (see §3.5.2), with those for our baseline dataset (see §3.3) in the second line. In all cases the reported "best fit" values correspond to the mode of the joint posterior distribution, and are denoted by the triangles in Figures 3.2-3.3. The uncertainties in the parameters are represented by the displayed 68% and 95% posterior contours, with the contours corresponding to the parameters of Eqn 3.2 evaluated with our baseline dataset colored blue.

3.5.1 Deterministic vs. Probabilistic M-R Relations

The primary motivation for this work was to assess the observational evidence for intrinsic scatter in the sub-Neptune M-R relation, and to characterize this scatter if

Table 3.2: Best-Fit Parameters of the M-R Relation

Equation	Dataset	C	γ	σ_M	β
1	RV only, < 4 $R_\oplus$	2.1	1.5	—	—
2	RV only, < 4 $R_\oplus$	2.7	1.3	1.9	—
2	dynamical TTVs only, < 4 $R_\oplus$	0.6	1.7	1.7	—
2	Weiss (< 4 $R_\oplus$)	2.8	0.9	2.5	—
2	RV only, < 1.6 $R_\oplus$	1.4	2.3	0.0	—
2	RV only, < 8 $R_\oplus$	1.6	1.8	2.9	—
3	RV only, < 4 $R_\oplus$	2.6	1.3	2.1	1.5

Note. — These "best fit" values correspond to the mode of the joint posterior distributions.

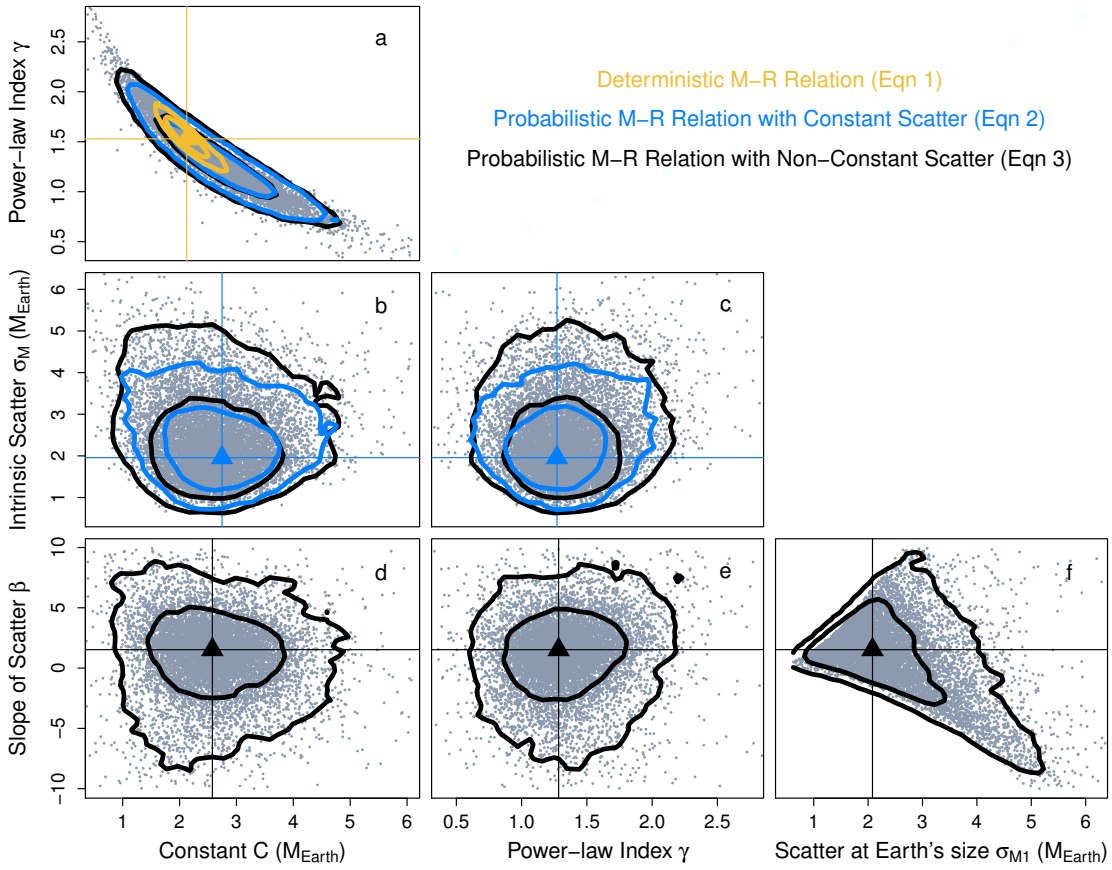

Figure 3.2: Posteriors for the parameters in our family of M-R relations (row 1: Equations 3.1-3.3; row 2: Equations 3.2-3.3; row 3 and gray posterior samples in all panels: Equation 3.3) when fit to our baseline dataset. 68% and 95% contours are shown for each, and demarcate the uncertainties on these M-R relation parameters; the triangles denote best-fit values. Panels b and c show that $\sigma_M = 0$ is strongly excluded for $R < 4$ $R_\oplus$, and so astrophysical scatter exists in the sub-Neptune M-R relation. Therefore, theoretical studies which require an M-R relation should use a probabilistic one like that of Eqn 3.2 with one of the sets of parameter values in Table 3.2.

warranted. To do so, we compare the posteriors for our three M-R relations in Figure 3.2 (note that not all relations have all parameters: for example, the deterministic M-R relation of Eqn 3.1 is described only by C and γ, so it only appears in panel a). Panels b and c show that this intrinsic scatter exists: because the posteriors lie away from zero,

$\sigma_M = 0$ is strongly excluded by the data, even with the currently large individual mass error bars. This is not a result of our choice of priors: the parameterization in Eqn 5.9 is equivalent to $\sigma_M^2 \sim 1/\sigma_M^2$, which is the (uninformative) Jeffrey's prior for such scale parameters. This prior is strongly weighted toward zero, in contrast to the posterior we compute.

Comparing the different M-R relations, we see that the C, γ posterior for the model given by Eqn 3.1 is much tighter than that for Eqns 3.2-3.3. This is expected: when we keep the dataset fixed but add more parameters, especially one like σ_M that by construction allows wiggle room around a deterministic relation, the observational information content per parameter decreases, and the posteriors widen. Given this expectation, what is arguably more notable are the small differences between Eqn 3.2 and 3.3's model posteriors for the parameters they have in common: most of the extra width of Eqn 3.3's joint posterior is contained in the new parameter β (Figure 3.2, panels d-f), which spans zero. There is therefore not enough evidence in the current dataset to justify an intrinsic scatter that changes as a function of radius in the way that we have parameterized it[2].

3.5.2 Changing the Dataset

The results in §3.5.1 are for our baseline dataset, an RV-only sample with $R_{obs} < 4$ $R_\oplus$. However, all Bayesian results depend on the data that are used, so it is important to carefully consider what the dataset contains. To demonstrate this, we

[2]While outside the scope of this work, future analyses of the M-R relation can address this and other questions of model selection more quantitatively by computing posterior Bayes factors. Regardless, the results for the statistical models represented by Eqns 3.1 and 3.3 can serve as a sensitivity test for that of Eqn 3.2, as we describe.

present some illustrative examples of the M-R relation posteriors under different mass and radius selection functions (Figure 3.3).

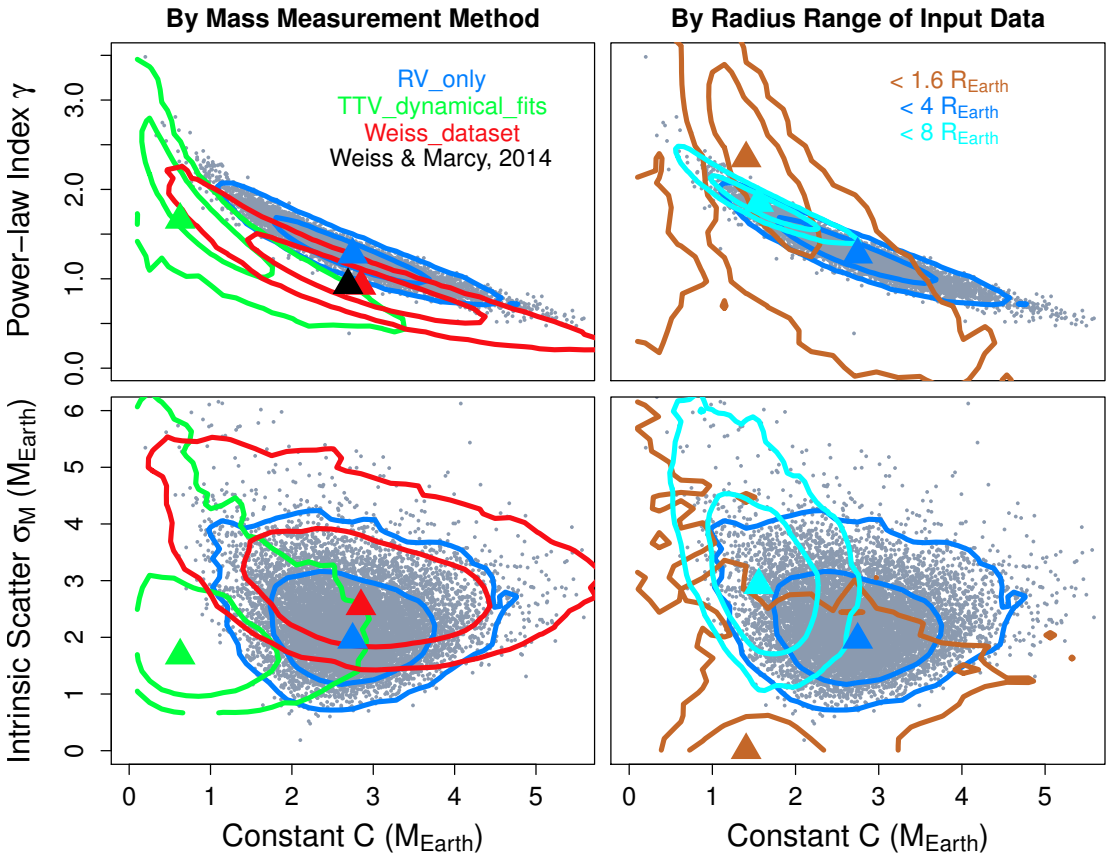

Figure 3.3: Posteriors for Eqn 3.2's M-R relation parameters when we change the input dataset (68% and 95% contours shown; triangles are best-fit values). The blue contours represent the baseline dataset and are the same as those in panels a and b of Figure 3.2. The green TTV M-R relation is systematically shifted downward (lower C) compared to the default M-R relation, while the red WM14 dataset, a hybrid of the two, produces a posterior which falls between them (the black point is the WM14 result itself). When we consider different radius ranges, we see that $R_{obs} < 8$ R$_\oplus$ (cyan) produces a slightly down-shifted, steeper, and more dispersed M-R relation than the default $R_{obs} < 4$ R$_\oplus$ (lower C and higher γ, σ_M, although the posteriors do overlap), while the M-R relation for $R_{obs} < 1.6$ R$_\oplus$ (orange) is not well constrained (although $\sigma_M \approx 0$ for reasonable values of C).

The left side of Figure 3.3 displays results for samples of planets that have had their masses measured in different ways. A number of prior studies (e.g. Jontof-

Hutter et al. 2014, WM14) have noted that planets with high SNR TTVs tend to be systematically less dense than RV-detected planets. Our results confirm this: the green TTV-only posterior is shifted towards lower C with similar γ and σ_M, which produces on average lower masses for a given radius. Furthermore, the hybrid WM14 dataset yields the red posterior, which falls between the TTV-only and RV-only posteriors yet peaks at lower γ, illustrating that posterior modes (Bayesian "best fits") for joint datasets are not necessarily averages of the modes for separate subsets. This behavior can be understood when one considers that these TTV planets are preferentially larger than the RV planets: this bias pulls the joint M-R relation down at higher radii because the TTV planets there have lower masses (which lowers γ) but affects the relation at lower radii very little because there are few small TTV planets in our sample (which keeps C roughly the same).

The right side of Figure 3.3 displays results for samples of planets spanning different radius ranges, illustrating the effect that a somewhat arbitrary radius cut can have on one's results. Compared to our default sub-Neptune range, a $R_{obs} < 8$ $R_\oplus$ cut produces an M-R relation that is overall shifted down, is steeper, and has more intrinsic scatter (the cyan posterior has lower C and higher γ, σ_M). This is consistent with the Lissauer et al. (2011b) fit to Earth and Saturn over a similar radius range, although neither of these Solar System planets were included in our dataset. Meanwhile, the M-R relation is poorly constrained for the $R_{obs} < 1.6$ $R_\oplus$ sample, the radius range outside of which rocky planets likely do not occur (Rogers, 2015). This is because our $0 < M_t^{(i)} < \rho_{iron} * (R_t^{(i)})^3$ restriction is most severe for these small planets, allowing

only a small range of physically plausible masses. This range is completely spanned by most of the mass measurements (see right side of Figure 3.4), so there is little empirical extrasolar information for $R_{obs} < 1.2$ $R_\oplus$, and the orange posteriors are dominated by the few larger planets with well measured masses. With this sample, there is not currently enough observational evidence in this radius range to rule out a deterministic relation.

3.6 Discussion

3.6.1 Visualizing the M-R Relation

While the posterior contours in Figures 3.2-3.3 show the best-fit M-R relation parameters and their uncertainties, visualizing the M-R relation itself requires that they be mapped from parameter space to mass, radius space. There are at least two ways to do this with Bayesian analysis, and they are displayed in Figures 3.4-3.5.

First, one can simply take the best-fit values and plot the resulting relation, as was done in Figure 3.4. Here the 1σ width of the probabilistic relation, as parameterized by σ_M, is denoted by the faded colored region while the mean relation, as parameterized by C and γ, is the thick line of the same color. Note that the mean M-R relations extend into unphysical regimes for $R < 1$ $R_\oplus$; this is because the mass observations span the physically allowed region, as discussed in §3.5.2, leaving the M-R relation to be constrained primarily by the locations of the larger, higher mass planets. The presence of intrinsic scatter in our M-R relation nevertheless allows physically realistic masses to be assigned to the smallest planets; to force this requirement, we recommend

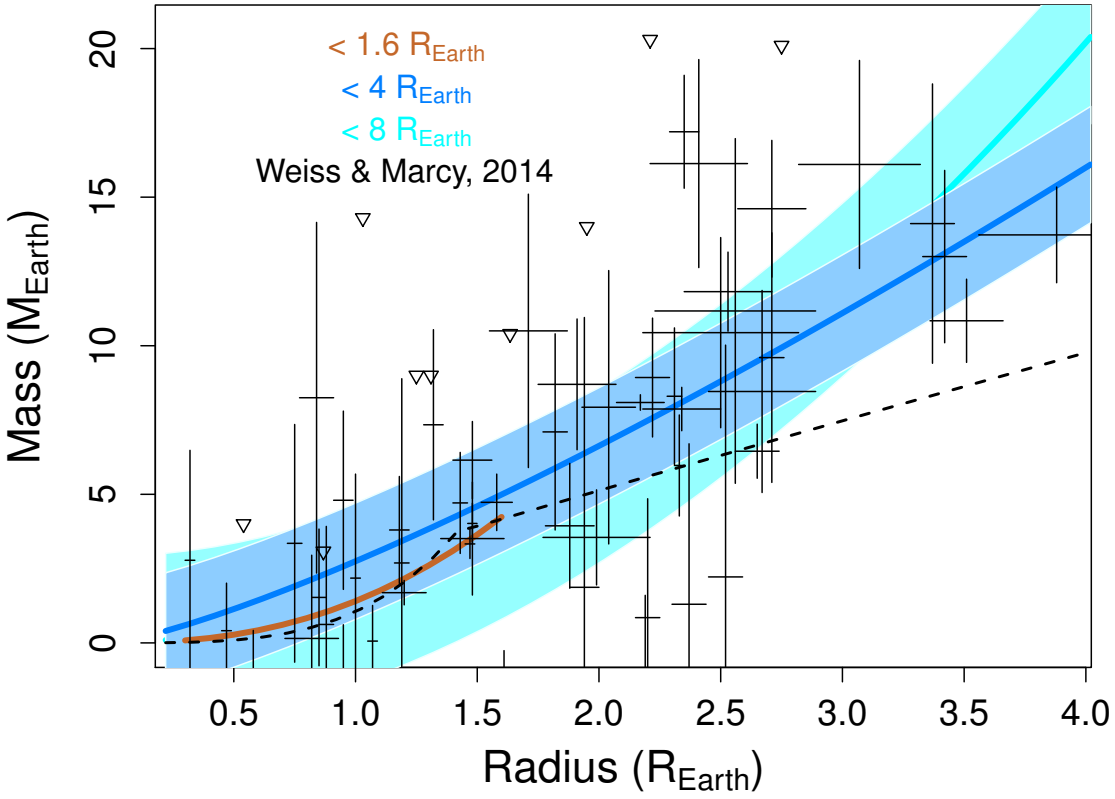

Figure 3.4: The best-fit M-R relations from the right column of Figure 3.3. For each, the solid line denotes the mean relation μ_M while the faded region denotes the standard deviation of the intrinsic scatter (vertical height of region to either side of line $= \sigma_M$; note $\sigma_M = 0$ for the smallest planets). The M-R relation of WM14 is the dashed black line while the baseline dataset is overplotted as the thin black lines with triangles for the upper limits (note that WM14 was calculated with a dataset that includes TTV planets).

adding a density constraint to Eqn 3.2 such that the probability of a planet being drawn outside this range is 0, or to use a different M-R relation for sub-Earth-sized planets. The different colors in the left panel correspond to the M-R relations in the right column of Figure 3.3; these mostly overlap in the sub-Neptune regime. Note that the RV-only dataset produces a steeper relation than one which also contains high SNR TTV planets (i.e. the black dashed WM14 relation), as discussed in §3.5.2.

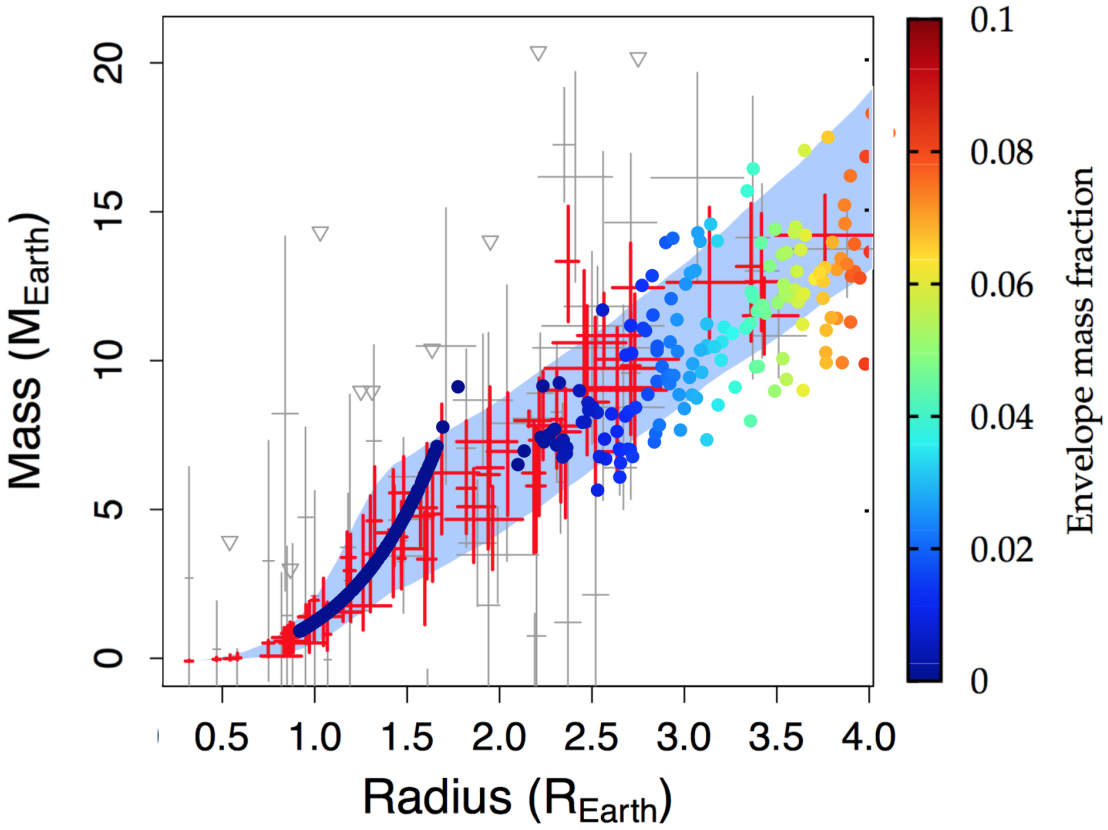

Figure 3.5: The default M-R relation (Eqn 3.2 with the baseline dataset) marginalized over the corresponding posterior distribution and subjected to our physical mass range restriction. The blue region now corresponds to the central 68% of planet masses that were drawn at a given radius. The posterior true masses and radii of individual planets are plotted red (the same R_{ob} and M_{ob} as in Figure 3.4 are plotted in gray for comparison). A sample of planets from the planet population synthesis studies of Jin et al. (2014) are overplotted, color-coded by the fraction of the planet's mass that is in a hydrogen and helium envelope.

While these best-fit M-R relations are easy to use, they do not take into account the fact that the posteriors have non-zero width and therefore a range of M-R relation parameters are allowed by any one dataset. A more thorough implementation of these results would incorporate these uncertainties by ranging over all of the posterior samples. This marginalization, which also incorporates the physical restrictions on M_t,

is displayed in Figure 3.5: now the blue region corresponds to the central 68% of planet masses that were drawn for a given radius. Note that this region is wider than that in Figure 3.4 and that the masses no longer extend into unphysical regimes. The posterior true masses and radii of individual planets in the baseline sample are plotted red, while the same R_{ob} and M_{ob} as on the left are plotted in gray. As expected[3], the posteriors have "shrunk" toward the mean relation within the uncertainties provided by the data.

Furthermore, a sample of planets from the planet population synthesis studies of Jin et al. (2014) are shown color-coded by the fraction of the planet's mass that is in a hydrogen and helium envelope. The qualitative agreement provided by this comparison motivates a more detailed, quantitative treatment of the characterization of the sub-Neptune population via the framework provided by theoretical modeling. This is the subject of Part II of this thesis; in particular, we use sophisticated statistical modeling to tie the observed planet radius distribution directly to the composition distribution of these planets, as parameterized by this same envelope mass fraction, f_{env}, in Chapter 5.

3.6.2 Using the M-R Relation to Predict Masses

The most straightforward and computationally simple way to map a sub-Neptune's radius to a mass while accounting for intrinsic scatter is to adopt Eqn 3.2 with one of the sets of parameters in Table 3.2 and impose a density constraint for the smallest planets. This best-fit M-R relation is analytic and represents a substantial

[3]Shrinkage is a well-known feature of hierarchical modeling, and is often desirable as it produces lower rms errors across the population than modeling individuals separately.

improvement over the previous deterministic relationships in capturing the full mass-radius distribution. However, it does not incorporate uncertainties in the M-R relation parameters or uncertainties in the measured planet radius itself. Depending on how detailed one's analysis needs to be, a more accurate predictive mass distribution may be needed.

To account for these issues, one must compute the posterior predictive distribution, which marginalizes over both the posteriors displayed here and the radius posterior produced by one's light curve modeling. This mass distribution will be wider than that produced by simply applying Eqn 3.2 (see right side of Figure 3.4) because it incorporates the above sources of uncertainty and thus more accurately reflects our state of knowledge about these planets' masses. Kepler-452 b (Jenkins et al., 2015) provides an example of an individual planet's posterior predictive mass distribution that has been calculated with this probabilistic M-R relation; because its computation requires the numerical posterior samples that we have produced, the resulting mass distribution is also numerical in nature. To enable more calculations like this one, we have posted our posterior samples in the github repository dawolfgang/MRrelation along with R code that uses them to compute and plot the posterior predictive mass distribution for individual planets.

3.7 Conclusions

In this chapter we have defined and constrained a probabilistic mass-radius relationship for sub-Neptune planets (Eqn 3.2 with parameter values in Table 3.2). In

particular, we demonstrate that there is intrinsic, astrophysical scatter in this relation, and that, except for the smallest planets, this scatter is nonzero for all considered datasets. For the first time in the exoplanet literature, we display the uncertainties in the M-R relation parameters through posterior distributions and explain how to properly incorporate these uncertainties into a predictive distribution of masses for individual planets. This M-R relation will be useful for anyone who wishes to perform large-scale dynamical or planet formation studies with the *Kepler* planet candidates.

Chapter 4

Adaptive Optics Follow-up of *Kepler*'s Sub-Neptunes

4.1 Introduction

As outlined in §1.2.3, acquiring follow-up observations of *Kepler* Objects of Interest (KOIs) is useful for a number of reasons, including the identification of likely false positives, i.e. transit signals which were dispositioned (see §1.2.1) as planet candidates but which are not actually planets transiting the specified target star. High-resolution imaging is especially valuable for such statistical validation of KOIs (as opposed to dynamical confirmation via the radial velocity method; see e.g. Torres et al. 2011; Morton 2012; Díaz et al. 2014), given the relative faintness of the typical planet candidate host star and the large on-sky spatial coverage of the *Kepler* pixels. Furthermore, the relatively small amount of telescope time and data processing that is required to iden-

tify possible contaminating sources compared to radial velocity follow-up makes high resolution imaging a particularly efficient observational technique for the validation of *Kepler* candidates as true planets. Even better, the sub-2″ resolution required for useful contributions to this effort is easily achieved with adaptive optics (AO) on 2-4m class telescopes, which facilitates a truly community-driven follow-up program which has the real potential to obtain the required telescope time to observe every identified KOI.

Helping to statistically validate KOIs is not the only application for measuring the fraction of KOIs that have close visual companions. When these sources, which may or may not be serendipitously projected unassociated background objects, are proven to be bound to the target star, the resulting fraction of planets present in multiple stellar systems can be used to constrain the range of environments in which typical planet formation processes occur, and to quantify the effect that stellar multiplicity has on these processes. Indeed, a number of theories even require the presence of a bound massive, distant companion to explain the existence of exoplanets with ∼day-long orbital periods, where full (both gas and solid) *in-situ* formation of these planets is particularly difficult. This class of planet formation and evolution theory includes planet-planet scattering (e.g. Rasio & Ford, 1996; Nagasawa & Ida, 2011), secular chaos (Wu & Lithwick, 2011), and secular evolution due to the long-term gravitational perturbations of an inclined distant companion (Holman et al., 1997; Wu & Murray, 2003; Fabrycky & Tremaine, 2007; Naoz et al., 2012; Batygin, 2012; Dawson & Chiang, 2014). While these are all reasonable proposals to explain close-in exoplanets, the question of how often they actually occur in nature remains. Comparing the fraction of short-period

planets that occur in multiple stellar systems to the fraction of those that don't can provide a highly valuable observational constraint on this question (further discussion of the theoretical application of observations like these is provided in §6.3).

4.1.1 Previous High-Resolution Imaging

Realizing the clear potential for high-impact contributions to the *Kepler* follow-up effort, a number of authors have undertaken high resolution imaging surveys of various subsets of *Kepler* Objects of Interest, including four major undertakings that involve > 100 targets, in addition to the one here. Other smaller samples include those of Wang et al. (2014), who characterized the stellar multiplicity of 56 KOIs with both radial velocities and near infrared (NIR) AO imaging at Keck, Palomar, and MMT for the purpose of comparing occurrence rates of planets in single and multiple stellar systems, and Everett et al. (2015) who obtained optical speckle imaging at Gemini North for 18 KOIs, along with a subset observed with NIR AO at Palomar and Keck, for the purpose of validating planetary candidates.

The first major effort is reported in Howell et al. (2011), who perform speckle imaging on 156 KOIs with the upgraded Differential Speckle Survey Instrument on the WIYN 3.5m telescope on Kitt Peak. This technique involves acquiring thousands of $\sim$ 30 ms exposures per object and combining them with significant post-processing to identify recurring binary patterns in the resulting speckles; the result is $\sim 0.2''$ resolution images of these KOIs with a typical limiting magnitude of $\Delta m = 4$ in three optical filters centered at 562, 692, 880 nm. With these data, they find 10 stars have detected additional sources within $1.4''$, the completeness limit of their FOV, yielding

a KOI visual companion fraction of 6%. However, their sample of KOIs were drawn from the earliest set of announced planet candidates (Borucki et al., 2011), which had a less mature vetting process and so included a number of signals later identified as false positives; correcting for this knowledge as of April 2014, this fraction drops to 3% (Lillo-Box et al., 2014). More recent results from this effort is presented in Horch et al. (2014); 49 companions were found around 623 stars, for a visual companion fraction of 8%. The host star sample has a *Kepler* apparent magnitude distribution that extends to $Kp \sim 15$, peaking at $Kp \sim 12.5$.

The second continuing effort is that of Lillo-Box et al. (2012) and Lillo-Box et al. (2014), who had obtained Sloan i-band "lucky imaging" of 234 KOIs as of April 2014 using the AstraLux instrument on the Calar Alto Observatory 2.2m telescope. This technique also involves taking thousands of very short exposures (100-200 ms) and combining the atmosphere-induced shifts during post-processing. As opposed to speckle imaging, however, only the top 10% of images with the best seeing, as quantified by the Strehl ratio (Strehl, 1902), are used, and the frames are combined by relatively simple centroid shifting. They are able to detect companions as close as $0.3''$ from the target star, and have typical limiting magnitudes of $\Delta m \sim 6 - 7$. Lillo-Box et al. (2014) find that, of the 174 planet candidate KOIs in their total sample, 117 are isolated (67.2%), 34 have at least one (visual) companion at separations of 3-6$''$ (20.1%), and 30 have companions closer than $3''$ (17.2%). Their KOI sample was selected on the basis of how interesting the planet candidates' properties were, and the distribution of the host stars' *Kepler* magnitudes peaks at $Kp \sim 13.5$, extending down to $Kp \sim 16$.

High-resolution imaging in the optical continued with (Law et al., 2014), who used Robo-AO, the laser guide star adaptive optics system on the fully robotic Palomar 60-inch telescope, to observe 715 KOIs in either the Sloan i-band or a special "LP600" filter which emulates the *Kepler* passband (Kp) between 600 and 900nm, with some additional transmission out to 1000nm; this cuts out the bluest end of the Kp filter. Consisting of by far the largest sample of high resolution follow-up images, these are the first results of an effort to uniformly observe every single KOI. For Law et al. (2014), they report that they randomly select KOIs from the Q1-6 planet candidate catalog (Batalha et al., 2013), which results in a host star apparent magnitude distribution that peaks at $Kp \sim 13.5$ and extends down to $Kp \sim 15.5$. Ranking the quality of their images as "low" ($\Delta m \sim 3$), "medium" ($\Delta m \sim 4$), and "high" ($\Delta m \sim 6$) with most images in the "medium" or "high" categories, they found 53 KOIs with visual companions between 0.2 and $2.5''$ (7%); when the target list is corrected to remove KOIs now known to be false positives, this becomes 49 out of 697 planet candidates (still 7%; Lillo-Box et al. 2014).

Finally, Adams et al. (2012), Adams et al. (2013), and Dressing et al. (2014) continue this effort in the NIR, the only large survey to do so, with natural guide star AO using ARIES on the 6.5m MMT and some earlier observations from PHARO on the Palomar Hale 200-inch telescope. They observed a total of 189 KOIs as part of the officially coordinated *Kepler* follow-up effort; imaging in J and K_s, they obtained an average spatial resolution of $\sim 0.2''$, with typical limiting magnitudes of $\Delta m \sim 5 - 6$ in K_s. Because they only use natural guide star AO, they are restricted to $Kp < 14$,

with their host star apparent brightness distribution peaking around $Kp \sim 12$. Across the three studies, they found that 32 KOIs (17%) had visual companions within $2''$; after false positives are removed and a different radius is considered to facilitate easier comparisons with the other surveys, this becomes 45 out of 165 KOIs (27%) having visual companions within $3''$.

Given the relative abundance of optical imaging among the large-scale *Kepler*'s high resolution follow-up campaigns, there is a valuable opportunity to contribute to the characterization of these planet candidates' stellar environments in the NIR. This is especially true for the faintest PC host stars that are out of the reach of natural guide star adaptive optics. Recognizing this potential for a high-impact contribution to this effort, I have acquired J, H, or K_s images for 196 KOIs with laser guide star AO on the Shane 3m telescope at Lick Observatory over the 2012, 2013, and 2014 observing seasons of the *Kepler* field.

4.2 Target Selection

My target selection strategy evolved from year to year as both as my science goals shifted focus and the above results were published; this was needed to provide the sample required for the analysis outlined in §4.6 and §6.3.2 while remaining complementary to the existing follow-up efforts detailed in §4.1.1. For the 2012 season, I focused on new KOIs from the Q1-6 catalog (Batalha et al., 2013) that had large ($\gtrsim 0.5''$) yet statistically insignificant ($< 3\sigma$) centroid offsets measured by the *Kepler* pipeline (this information was provided in the DV summaries; see §1.2.1). I also gave higher priority

to targets whose transit durations relative to their periods indicated either non-circular orbits or mischaracterized stellar radii (high impact parameters were ruled out upon visual inspection of the transit shape); this simultaneously identifies the KOIs most in need of follow-up and, if blends are ruled out and the stellar properties are further verified with spectroscopy, the KOIs that could have experienced a significant amount of post-formation dynamical evolution (e.g. Dawson & Johnson, 2012).

Starting with the 2013 season I began prioritizing the single-planet systems for two reasons. First, official *Kepler* follow-up efforts were focusing on the multiple-planet systems, an understandable choice given the richness of the information they provide; after all, several planets orbiting the same star produce additional constraints on the average stellar density and thus on their orbital eccentricities (Kipping et al., 2012), and they offer a window into the frequency of different modes of planet formation via the presence of mean motion resonances and the observed distribution of mutual inclinations (Lissauer et al., 2011b; Fabrycky et al., 2014). Interestingly, they also have a low *a priori* probability of being astrophysical false positives (Lissauer et al., 2014), a finding that has allowed the statistical validation of hundreds of multiple-system planetary candidates (Rowe et al., 2014); one could interpret this as as even stronger reason to follow up the single-planet systems that need follow-up observations the most to be statistically validated as bona fide planets.

Second, the population of single-planet systems also provides valuable insight into planet formation mechanisms, even if it is much less appreciated as such. As noted by Latham et al. (2011), the Jupiter-sized planets that *Kepler* has found tend to be

solitary, while the multiple-planet systems tend to be dominated by sub-Neptune-sized planets. At the same time, *Kepler* has found an enormous number of single Neptune-sized planet candidates — over 2200 with $R < 6$ $R_\oplus$ as of the Q1-17 DR24 planet candidate catalog (Akeson et al., 2013) — that cannot be explained by invoking *Kepler*'s detection biases with a well-behaved inclination distribution (Hansen & Murray, 2013). This hints suggestively at a different formation scenario for these single-planet systems: rather than these planets reaching their current small spatial distance from their host stars by slow, orderly migration through a protoplanetary disk, these planets instead could have come to their current locations by Kozai-Lidov oscillations (Kozai, 1962; Lidov, 1962), a secular process involving a hierarchical triple system where eccentricity and inclination of the inner planetary orbit oscillate back and forth to extreme values. When the eccentricity becomes large enough that the planets periapse enters the regime where tidal interactions with the star become important, the planets orbital energy will quickly dissipate, leaving the planet in a very compact orbit around its host star (Holman et al., 1997; Fabrycky & Tremaine, 2007; Naoz et al., 2012). While there are other dynamical evolution mechanisms which qualitatively produce this same result, only Kozai-Lidov oscillations require that the gravitational perturber still be present and bound to the host star. Fortunately, this constraint is testable by observations, and in particular high resolution images like the ones we obtain here. Extending observations into the NIR is especially important, as it is more sensitive to low-mass stellar companions compared to the optical bandpasses used by Robo-AO or the lucky imaging technique.

The promise of using the large sample of single-Neptune-sized planetary systems to test the importance of the Kozai mechanism for planetary orbital evolution (see §6.3.2) motivates the rest of our target selection. In particular, we prioritized KOIs with $R < 6$ $R_\oplus$ that were the only observed planet candidate in their system. Given that there were over 1700 of these by the start of the 2013 observing season, we narrowed our target list further by choosing two subsets of planets: those with signal-to-noise ratios $\gtrsim 30 - 50$, which have the highest potential for being statistically validated (cf. Díaz et al., 2014), and those with very long or very short transit durations, which may be eccentric and could represent a class of "failed Kozai" planets (Dawson & Johnson, 2012; Dawson & Chiang, 2014). We also observed some of the KOIs that Robo-AO discovered to have close companions to provide color information that could constrain their spectral types and assist efforts to determine whether or not these additional sources are gravitationally bound to the target star. We also include some of the fainter multiple-planet KOIs to provide a control sample that may not yet have been observed in the NIR.

4.3 Observations

Given our goal to provide a large, complementary sample of NIR follow-up observations, we needed a facility that can provide both a significant amount of observing time over an extended period and a laser guide star-enabled adaptive optics (LGS AO) system to facilitate imaging objects dimmer than 14th magnitude, which constitutes most of *Kepler*'s single Neptune KOIs (see Figure 4.1). Lick Observatory's Shane 3m

telescope meets both of these requirements; furthermore, as a UC graduate student I could P.I. the proposals, thereby lowering the barrier to entry on this effort.

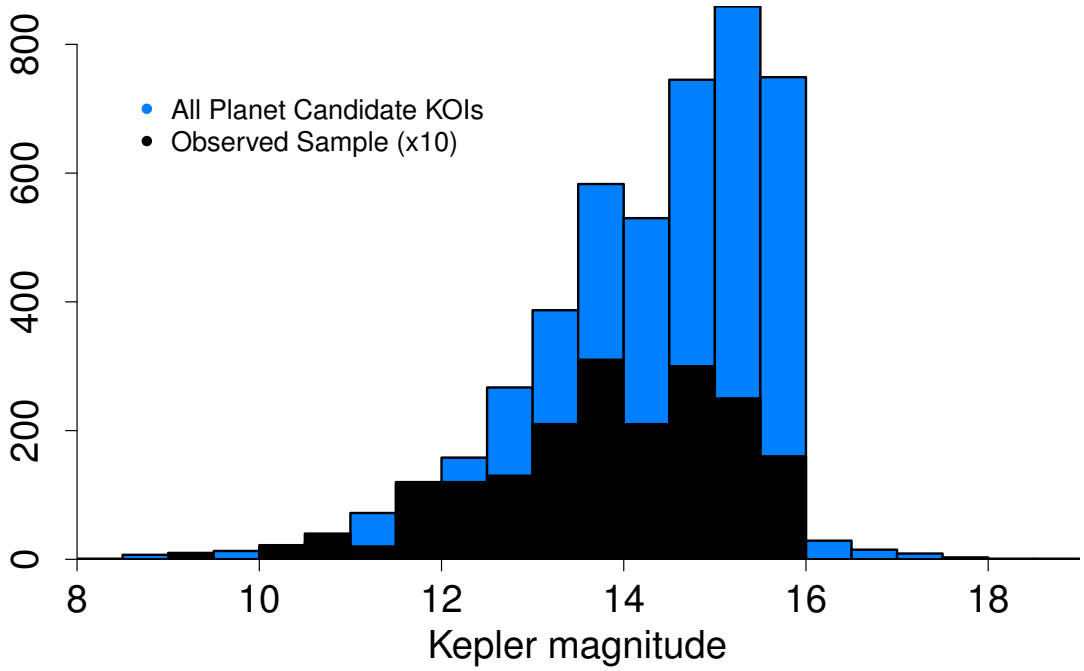

Figure 4.1: *Kepler* magnitude distribution of our AO sample versus all planet candidates in the Exoplanet Archive cumulative table as of the Q1-17 DR24 release (the counts in our sample have been inflated by a factor of 10 to facilitate easier comparison). Note the focus of our observing strategy on KOIs with $13.5 < Kp < 16$.

During the 2012 and 2013 observing seasons, we acquired high resolution images with the Lick LGS AO system (Brase et al., 1994; Max et al., 1997) and IRCAL, the associated NIR camera (Lloyd et al., 2000). For KOIs with $r' < 13.5$ we operated the system in natural guide star (NGS) mode and guided off the target itself, which is possible when the star provides enough signal on the wavefront sensor to measure the atmospheric distortion along its own line of sight. LGS mode enabled observations of the dimmer KOIs, for which high-order corrections are infeasible without an additional

light source; however, this mode still requires low-order correction be provided by a "tip-tilt" $r' < 16$ star. Fortunately, the vast majority of KOIs are above this magnitude limit, meaning that we have on-axis correction for all of our targets, thereby achieving the highest Strehl ratios that are possible with the system (0.4 in K_s with good natural seeing conditions; Olivier et al. 1994; Srinath et al. 2014).

In practice we performed most of our imaging in H on IRCAL, as that bandpass best balanced the low sky background of the shorter NIR wavelengths with the higher Strehl ratios possible in the longer wavelengths. Along with 1-2$''$ native seeing, we obtained corrections that on average corresponded to Strehl ratios of ~ 0.2. While this is fairly low for modern AO systems, a pixel scale of $\sim 0.075''$ and a typical corrected FWHM of ~ 5 pixels nevertheless allowed us to resolve companions as close as $0.5''$. Furthermore, the field of view of the 256 x 256-pixel detector was $20''$ x $19''$, allowing detections of companions out to $> 15''$ in our combined dithered images; this spans the area covered by a typical *Kepler* "optimal" aperture (6-10$''$ across) from which a given target star's photometric time series is calculated.

As the Lick LGS AO system approached its 20th anniversary, it experienced a major upgrade. The new ShaneAO system (Gavel et al., 2014) with ShARCS, a camera with grism spectroscopy and a 2k x 2k HAWAII-2RG detector (McGurk et al., 2014), produces higher Strehl ratios (up to 0.8 in K_s with a new laser currently under development; up to 0.6 for our observations) and a full magnitude lower sky background (Srinath et al., 2014). In practice, this has allows us to achieve the same $\Delta m \sim 5 - 7$ between our target and the detected companions in a factor of 2-3 less time, including

the overhead involved in reading out the much larger detector and the slightly longer acquisition time between telescope nods. Accordingly, we were able to triple our sample size during the 2014 season alone. Furthermore, the new system allows us to achieve higher resolutions; with typical $1 - 1.5''$ native seeing we are able to clearly resolve the $0.25''$ binaries detected by Robo-AO and lucky imaging.

In total, we were awarded 9 nights of observing time on IRCAL in August 2012, 5 nights in June 2013, 8 nights of ShARCS shared-risk observing in May and July 2014, and 4 full nights and 5 half-nights in August-October 2014. Factoring in weather and engineering-associated losses, we obtained the equivalent of 23 nights of good observing, with 21 of those dedicated to imaging follow-up of *Kepler* planet host stars. We performed most of our IRCAL imaging in H, as explained above, and our ShARCS observing in K_s, given the excellent correction and much lower sky background provided by the new instrument. For a few select targets, usually those for which a close companion was immediately visible upon read-out of the detector, we obtained either K_s (IRCAL) or J (ShARCS) images for color information that can constrain the spectral type of the detected source; along with the measured apparent magnitude, this can help determine whether the visual companion is at the same distance from us as the target star. We also performed standard 5-point dithering for each target with $4 - 5''$ offsets to accurately characterize and then subtract the sky background in the vicinity of the target.

4.4 Data Reduction

For both the IRCAL and ShARCS data we used a modified version of an IDL data reduction pipeline that was originally written to reduce and combine dithered Keck NIRC2 images (Rosalie McGurk, private communication). The pipeline performs standard dark subtraction and flat fielding, computes and subtracts a sky background image using the extra area covered by the frame-to-frame dithers, corrects for bad pixels with an optional bad pixel map, and combines the images with subpixel centroiding that is guided by interactive source identification from the user.

Once reduced, the images are processed with PyRAF to identify and perform photometry on all stars in the field of view. In particular, we use the *imexamine* task to identify sources by eye and to calculate initial centroid positions, then the *phot* and *psf* routines of the PyRAF-bundled DAOPHOT package (Stetson, 1987) to perform aperture photometry and point spread function (PSF) fitting, respectively. For close binaries, the PSF fitting is performed iteratively: the dimmer star is subtracted from the image and the PSF that is fit to the brighter star is recomputed until its parameters converge to a single set of values. We fit the data with an analytic PSF only, using no look-up tables (see discussion below); specifically, for IRCAL we allow the *psf* task to choose which profile best fits the data (usually a Gaussian core with Lorentzian wings), but due to computational concerns force the ShARCS data to be fit to a Lorentzian. Both of these functions are decent approximations for the observed PSFs, as adaptive optics by design produces a sharp central core with broad, dim wings. In practice our PSF fitting leaves both positive and negative residual structure that, pixel-by-pixel, is

usually $\sim 10\%$ of the original photon counts.

Because DAOPHOT was developed for crowded field photometry, it is not the ideal software package for measuring the magnitudes of stars in AO images that have been dithered and combined. In particular, the point spread function of an AO image is not guaranteed to be constant across the entire field of view due to the less accurate real-time measurement of atmospheric turbulence that is off-axis from the guide star (this effect is known as anisoplanatism). Additionally, there can be non-negligible variation of the PSF from frame to frame, so that if a star falls outside of our $5''$ dithering box, the combined PSF may be different of that from our target. With these caveats in mind, we make no attempt to perform absolute photometry on our images, limit ourselves only to the analytic PSF fitting described above, and offer the relative magnitudes in Table 4.1 with typical errors of $\pm$ 0.1 mag.

4.5 Results

In short, 42 KOIs in our sample (21%) have visual companions within $3''$, and 90 (46%) have them within $6''$. These proportions are consistent with the results from the deepest high resolution follow-up surveys detailed in §4.1.1. The separations and relative magnitudes of every source that has been detected in our AO images are provided in Figure 4.3; the details for those within $10''$ of the KOI in question, which are sources that could have contaminated the *Kepler* photometry given the typical 6-$10''$ aperture size, are listed in Table 4.1. The KOIs which we did not observe to have visual companions within $10''$ are listed in Table 4.2. We emphasize that we do not expect

all, or even most, of these sources to be bound to the KOIs; this is especially true for the farthest sources. Fortunately, the color information that we have obtained as well as optical-NIR colors that are available when the KOI was observed by other studies can help constrain their radial distances. This endeavor, which is part of the effort to quantify the probability of companionship, is outlined in §4.6.

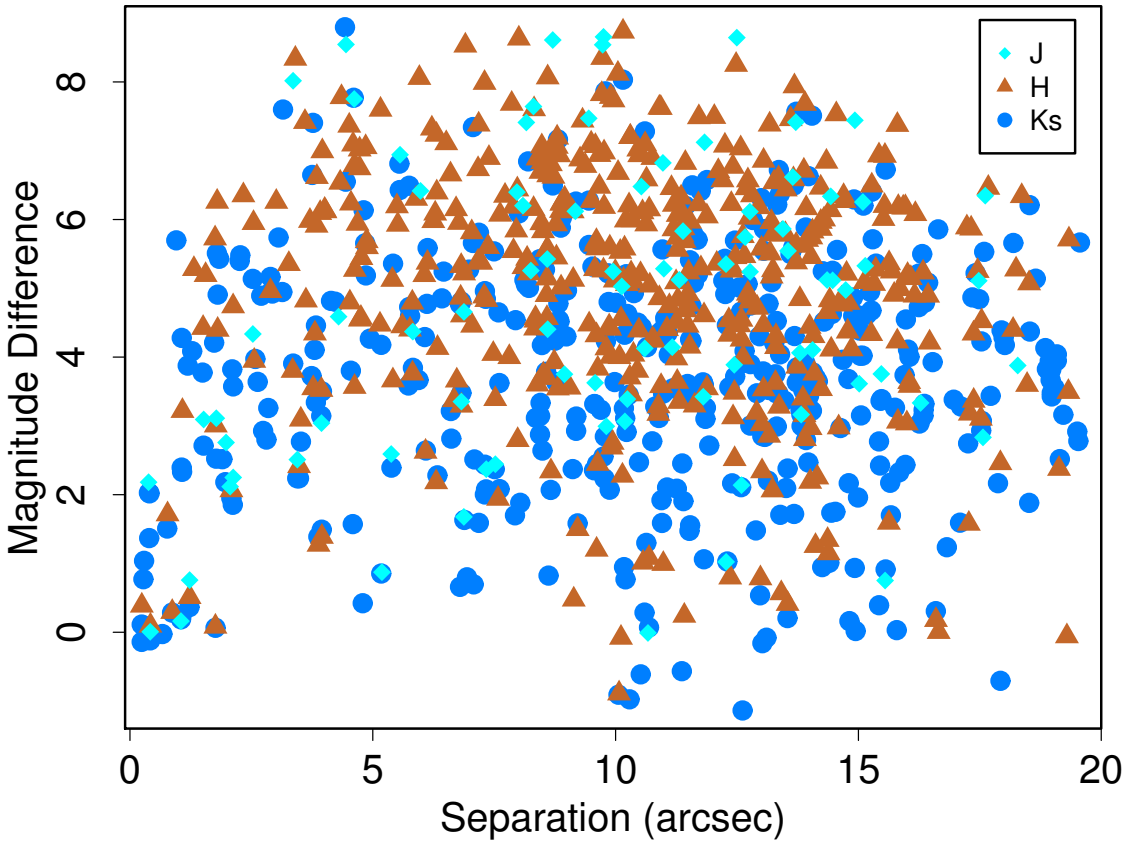

Figure 4.2: Magnitude difference vs. separation between each KOI in our sample and the additional sources detected in our AO images, denoted by the associated observing filter. 42 KOIs in our sample (21%) have visual companions within 3″, and 90 (46%) have them within 6″; the details of those within 10″ are listed in Table 4.1.

Table 4.1: Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
49	9527334	130604	H	9.62	75.8	5.58
70	6850504	120828	H	3.81	54.0	4.34
			Ks			4.10
103	2444412	130604	H	3.96	47.3	6.99
103	2444412	130604	H	9.76	278.9	6.18
119	9471974	141006	J	1.05	118.0	0.16
		141005,141006	Ks			0.22
119	9471974	141006	J	8.09	269.5	6.20
119	9471974	141006	J	9.58	210.1	3.63
		141005,141006	Ks			3.32
162	8107380	140819	Ks	0.28	123.7	0.10
162	8107380	130605	H	3.27	356.6	5.35
		140819	Ks			4.95
162	8107380	130605	H	6.06	200.9	5.19
		140819	Ks			4.77
162	8107380	130605	H	7.37	127.1	5.60
162	8107380	130605	H	7.95	198.9	5.84
165	9527915	130603	H	4.52	98.4	7.36
165	9527915	130603	H	7.39	345.9	6.90
165	9527915	130603	H	7.72	242.4	6.35

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
177	6803202	141005	Ks	0.24	218.8	0.11
177	6803202	141005	Ks	6.61	108.4	3.22
240	8026752	140818	Ks	2.53	4.4	5.14
240	8026752	140818	Ks	2.74	270.8	2.93
242	3642741	140818	Ks	4.55	128.6	3.80
242	3642741	140818	Ks	9.12	106.8	2.37
242	3642741	140818	Ks	9.80	65.9	4.23
257	5514383	140816	J	4.45	113.4	8.54
			Ks			8.80
268	3425851	140818	J	1.78	265.2	3.11
		130602	H			3.01
		140818	Ks			2.52
268	3425851	140818	J	2.53	307.4	4.34
		130602	H			5.95
268	3425851	140818	J	9.75	276.8	8.65
		130602	H			7.82
		140818	Ks			7.86
284	6021275	120801	H	0.87	96.6	0.29
			Ks			0.29
284	6021275	120801	H	5.96	174.0	8.06

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
305	6063220	140717	Ks	5.56	94.4	6.43
306	6071903	140816	J	2.13	244.1	2.25
			Ks			1.85
306	6071903	140816	J	4.62	325.2	7.75
			Ks			7.77
306	6071903	140816	J	5.56	139.2	6.94
			Ks			6.81
306	6071903	140816	J	9.17	255.7	6.12
			Ks			6.26
306	6071903	140816	J	9.45	48.5	7.47
346	11100383	130601	H	1.56	351.7	5.19
346	11100383	130601	H	7.18	59.2	7.38
346	11100383	130601	H	7.30	17.9	7.99
355	11621223	130603	H	4.59	304.3	7.09
355	11621223	130603	H	6.25	296.7	5.98
355	11621223	130603	H	6.67	325.4	3.37
361	12404954	141007	Ks	8.57	104.8	3.79
429	10616679	140819	Ks	6.46	142.9	5.24
432	10858832	140817	Ks	9.70	202.2	3.23
480	11134879	140718	Ks	4.86	1.7	5.19

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
531	10395543	130601	H	8.31	169.7	6.15
531	10395543	130601	H	9.99	59.5	5.24
535	10873260	140819	Ks	6.62	172.7	2.82
535	10873260	140819	Ks	7.31	176.9	2.44
578	8565266	140910	Ks	2.93	240.9	4.92
578	8565266	140910	Ks	6.98	339.6	5.27
578	8565266	140910	Ks	8.45	127.9	3.33
618	10353968	140708	Ks	7.08	224.3	0.70
618	10353968	140708	Ks	8.36	217.0	3.12
640	5121511	140708	J	0.42	300.6	-0.01
		130604	H			0.10
		130604,140708	Ks			-0.12
640	5121511	130604	H	5.77	24.3	4.40
		130604,140708	Ks			4.60
640	5121511	140708	J	5.82	25.3	4.37
640	5121511	140708	J	8.26	196.6	5.26
		130604	H			5.31
		130604	Ks			5.05
640	5121511	130604	H	8.58	2.4	5.56
			Ks			5.42

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
652	5796675	130602	J	1.23	272.4	0.76
			H			0.51
			Ks			0.37
652	5796675	130602	J	6.82	351.0	3.35
			H			3.29
			Ks			3.48
652	5796675	130602	J	8.59	187.5	5.43
			H			5.14
			Ks			5.27
666	6707835	130604	H	7.48	356.8	4.05
666	6707835	130604	H	8.59	227.4	4.57
672*	7115784	130604	H	9.05	105.6	3.56
687	7976520	140818	Ks	8.04	88.8	1.88
697	8878187	140708	Ks	0.67	53.8	-0.03
697	8878187	140708	Ks	7.31	94.5	2.00
714	9702072	130603	H	8.34	163.4	4.41
714	9702072	130603	H	9.61	31.3	4.32
746	10526549	140718	Ks	4.23	2.8	4.80
747	10583066	130605	H	4.02	176.3	6.12
747	10583066	130605	H	8.95	283.5	6.31

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
769	11460018	140718	Ks	1.19	166.0	3.88
769	11460018	140718	Ks	6.94	336.6	0.80
769	11460018	140718	Ks	8.49	206.9	2.64
826	5272878	140816	Ks	1.07	205.9	4.28
843	5881688	140816	Ks	7.93	8.1	1.70
843	5881688	140816	Ks	8.75	28.7	4.24
875	7135852	130601	H	1.32	278.7	5.28
		140718	Ks			4.09
875	7135852	130601	H	5.15	221.2	4.47
		140718	Ks			4.18
875	7135852	130601	H	8.77	64.3	5.93
914	8552202	140817	Ks	4.79	183.9	0.42
922	8826878	141006	Ks	2.70	221.6	4.89
922	8826878	141006	Ks	8.44	133.7	2.88
923	8883593	140817	Ks	4.59	312.1	1.57
923	8883593	140817	Ks	8.45	301.6	3.14
984	1161345	130601	H	1.76	223.1	0.08
			Ks			0.06
984	1161345	130601	H	4.85	144.1	7.05
987	7295235	130601	Ks	1.96	227.0	2.18

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
987	7295235	130601	Ks	8.21	3.3	5.01
988	2302548	130601	H	7.17	186.0	6.73
988	2302548	130601	H	9.13	179.7	5.12
1002	1865042	140818	Ks	0.29	173.7	1.04
1109*	3235672	120830	H	3.84	56.7	5.91
1109*	3235672	120830	H	4.77	302.9	5.44
1109*	3235672	120830	H	8.35	194.2	4.53
1116	2849805	130605	H	5.89	40.2	6.30
1116	2849805	130605	H	6.88	196.8	5.22
1116	2849805	130605	H	8.50	252.4	5.83
1150	8278371	140718	Ks	0.40	320.7	1.37
1150	8278371	130605	H	7.17	167.0	5.39
1274	8800954	130601	H	1.08	243.5	3.22
		130601,140708	Ks		242.6	2.39
1315	10928043	130601	Ks	1.79	26.2	5.51
1357	6719086	140816	Ks	2.85	84.4	3.26
1357	6719086	140816	Ks	3.83	164.7	3.34
1357	6719086	140816	Ks	9.75	124.6	2.55
1397	9427402	140818	Ks	6.42	41.7	4.85
1397	9427402	140818	Ks	8.62	44.0	0.82

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
1425	11254382	141006	J	5.19	342.4	0.87
			Ks			0.85
1425	11254382	141006	J	7.34	275.5	2.37
			Ks			2.08
1428	11401182	140817	Ks	2.63	168.7	3.64
1442	11600889	130604	H	2.12	71.4	4.74
		130604,140708	Ks			3.82
1442	11600889	130604	H	9.76	270.6	6.97
1481	9597806	141006	Ks	4.14	170.8	4.82
1515	7871954	130601	H	9.38	154.6	6.78
1588	5617854	140816	Ks	4.84	169.7	5.65
1597	5039228	140817	Ks	4.44	245.0	6.55
1597	5039228	140817	Ks	5.75	329.5	6.49
1597	5039228	140817	Ks	7.59	138.6	4.65
1597	5039228	140817	Ks	8.81	309.2	7.16
1606	9886661	130604	H	3.62	100.5	4.82
1606	9886661	130604	H	6.33	166.9	4.13
1606	9886661	130604	H	9.63	286.4	3.83
1615	4278221	130604	Ks	4.81	266.3	6.13
1615	4278221	130604	Ks	7.50	172.1	5.54

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
1615	4278221	130604	Ks	8.12	168.5	5.12
1615	4278221	130604	Ks	9.45	253.2	5.63
1615	4278221	130604	Ks	9.98	77.4	3.64
1619	4276716	140526	J	2.06	224.3	2.12
			H			2.06
			Ks			1.95
1626	6387542	120831	H	9.14	154.6	0.48
1639	10749128	120829	H	8.58	250.8	4.90
1665	4932442	120807	H	5.81	12.1	4.79
1665	4932442	120807	H	8.87	36.2	6.12
1665	4932442	120807	H	9.27	238.4	6.87
1665	4932442	120807	H	9.94	31.4	2.75
1701	7222086	140819	Ks	7.06	248.8	7.34
1701	7222086	140819	Ks	8.20	153.1	6.84
1702	7304449	141005	Ks	8.44	200.4	5.29
1747	7032421	120828	H	9.98	220.8	5.11
1751	9729691	140819	Ks	2.89	196.9	5.16
1751	9729691	140819	Ks	5.73	30.2	3.59
1781	11551692	141007	J	3.45	329.7	2.51
		120802	H			2.42

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
		141007	Ks			2.24
1781	11551692	120802	Ks	3.48	329.9	2.24
1784	10158418	140819	Ks	0.28	287.5	0.77
1786	3128793	120830	H	9.75	245.7	5.26
1802	11298298	140819	Ks	5.42	238.0	5.36
1821	8832512	140818	Ks	3.83	88.1	4.46
1835	9471268	141007	Ks	7.09	148.0	2.51
1838	5526527	130601	H	4.73	180.6	4.54
1838	5526527	130601	H	7.81	182.3	4.00
1843	5080636	141006	Ks	3.15	263.4	7.60
1849	9735426	140817	Ks	9.19	71.4	3.14
1858	8160953	120802	H	6.72	289.4	6.15
1860	4157325	120802	H	7.04	255.3	5.81
			Ks			5.66
1860	4157325	120802	Ks	7.19	136.0	5.81
1860	4157325	120802	H	7.34	266.0	4.83
			Ks			4.85
1860	4157325	120802	H	8.48	227.7	7.06
1860	4157325	120802	H	9.32	237.4	7.44
			Ks			5.32

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
1860	4157325	120802	H	9.88	91.8	5.59
			Ks			5.22
1868	6773862	130601	H	4.88	356.5	5.60
1868	6773862	130601	H	8.35	4.8	4.57
1877	10454632	130605	H	3.60	152.0	7.42
1877	10454632	130605	H	6.51	357.4	7.11
1877	10454632	130605	H	8.61	81.5	6.64
1880	10332883	140718	Ks	1.73	97.2	4.21
1886	9549648	140819	Ks	4.94	249.6	4.27
1886	9549648	140819	Ks	8.48	28.8	4.47
1890	7449136	140816	Ks	0.40	142.3	2.02
1890	7449136	140816	J	7.97	14.9	6.40
			Ks			6.09
1890	7449136	140816	J	8.16	269.1	7.41
1904	8766650	130603	H	7.90	289.3	5.54
1904	8766650	130603	H	8.88	282.3	5.01
1915	9101496	141006	Ks	0.95	272.9	5.70
1916	6037581	120830	H	6.77	302.8	5.95
1937	10190777	130603	H	2.34	243.9	6.35
1957	10028352	140817,140819	Ks	2.26	83.2	5.40

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
1957	10028352	140817,140819	Ks	8.83	74.3	4.54
1957	10028352	140817	Ks	9.45	61.8	6.27
1964	7887791	140718	J	0.38	359.9	2.18
			Ks			2.03
1973	4917596	130605	H	0.77	24.1	1.71
			Ks			1.51
1973	4917596	130605	H	4.33	47.0	6.54
1973	4917596	130605	H	4.73	153.0	6.75
1973	4917596	130605	H	8.47	282.0	5.95
1973	4917596	130605	H	9.81	223.4	4.13
1985	8142942	140718	Ks	2.81	152.3	2.80
1985	8142942	140718	Ks	8.55	126.5	4.37
1988	9044228	140819	Ks	9.00	335.4	4.97
2009	2449431	140818	J	1.52	175.3	3.09
			Ks			2.71
2020	9349482	120830	H	8.59	339.6	3.70
2034	3657758	140718	Ks	5.94	341.3	3.67
2034	3657758	140718	Ks	7.18	212.1	1.59
2034	3657758	140718	Ks	9.19	285.6	2.93
2036	6382217	120806	H	6.31	204.3	7.23

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
2036	6382217	120806	H	6.82	110.0	5.32
2036	6382217	120806	H	7.29	87.1	4.88
			Ks			4.96
2036	6382217	120806	H	8.45	231.4	6.69
2036	6382217	120806	H	9.51	75.3	6.27
2036	6382217	120806	Ks	9.55	43.8	2.36
2036	6382217	120806	H	9.58	279.2	5.28
2036	6382217	120806	H	9.67	343.5	6.94
2038	8950568	120801	H	1.80	308.7	6.26
2038	8950568	120801	H	4.55	175.5	6.23
2038	8950568	120801	H	7.62	258.6	6.07
2038	8950568	120801	H	8.39	294.9	3.60
2038	8950568	120801	H	9.67	94.8	6.48
2038	8950568	120801	H	9.92	183.8	2.68
2169	9006186	120806	H	3.52	68.3	3.09
			Ks			2.77
2174	8261920	120806	H	1.51	238.0	4.42
			Ks			3.77
2174	8261920	120806	Ks	3.06	198.3	5.74
2174	8261920	120806	H	3.88	132.4	1.27

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
			Ks			1.39
2174	8261920	120806	H	3.95	145.1	1.39
			Ks			1.49
2174	8261920	120806	H	4.62	41.1	6.79
2174	8261920	120806	H	5.50	292.0	6.19
2174	8261920	120806	H	8.78	172.5	5.45
			Ks			5.88
2174	8261920	120806	H	9.22	106.7	1.50
			Ks			1.58
2174	8261920	120806	H	9.50	140.4	7.10
2174	8261920	120806	H	9.58	81.2	7.07
2215	7050060	120831	H	3.92	147.6	6.09
2215	7050060	120831	H	4.76	265.4	4.57
2215	7050060	120831	H	6.31	66.0	2.18
			Ks			2.28
2215	7050060	120831	H	9.46	330.9	4.47
2215	7050060	120831	Ks	9.88	280.9	2.07
2215	7050060	120831	H	9.90	175.2	4.00
2219	5357545	130603	H	4.53	237.8	4.79
2219	5357545	130603	H	5.41	296.3	3.66

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
2219	5357545	130603	H	6.73	224.9	3.66
2219	5357545	130603	H	8.72	211.8	6.62
2224	8892157	120830	H	8.18	246.1	5.43
2248	11030475	120801	H	3.94	198.4	3.52
			Ks			3.50
2248	11030475	120801	H	6.09	82.2	2.62
			Ks			2.64
2248	11030475	120801	H	6.29	66.0	6.09
2248	11030475	120801	H	8.51	169.0	3.75
			Ks			4.18
2248	11030475	120801	H	8.81	167.9	3.54
			Ks			3.64
2311	4247991	120802	H	6.91	46.9	8.53
2311	4247991	120802,120803	H	8.79	309.1	7.06
2311	4247991	120802,120803	H	8.88	324.9	6.23
		120802	Ks			6.02
2324	7746958	140708	J	5.38	267.8	2.59
			Ks			2.39
2324	7746958	140708	J	6.88	13.6	4.66
			Ks			4.81

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
2410	8676038	120807	H	4.69	330.3	5.94
2470	9108085	120828	H	3.01	178.1	6.26
2470	9108085	120828	H	9.58	229.9	5.63
2477	6846911	120830	H	4.63	83.7	3.57
2486	8074328	141005	Ks	0.24	250.4	-0.14
2486	8074328	141005	Ks	6.07	83.9	4.29
2519	4047631	120828	H	2.89	345.1	4.96
		120831	Ks			4.95
2519	4047631	120828	H	6.32	211.9	6.39
2519	4047631	120828	H	8.66	290.4	7.01
2519	4047631	120828	H	9.63	178.5	2.45
		120831	Ks			2.45
2521	7183745	120828	H	7.58	123.9	1.94
			Ks			2.08
2522	9177629	120828	H	7.86	290.2	7.68
2587	5546691	120831	H	9.61	139.6	1.20
2650	8890150	120828	H	6.13	10.6	5.37
			Ks			5.58
2678	6779260	130602	J	3.36	25.2	8.02
			H			8.34

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
2678	6779260	130602	J	8.31	147.0	7.64
			H			7.60
2678	6779260	130602	J	8.71	259.9	8.61
			H			8.07
2678	6779260	130602	J	9.73	100.2	8.54
			H			8.35
2699	6690836	140817	J	3.94	291.9	3.04
			Ks			3.14
2700	8639908	140708	Ks	6.80	303.3	0.66
2705	11453592	140910	J	1.98	302.8	2.76
			Ks			2.52
2705	11453592	140910	J	5.98	290.9	6.42
2707	5480640	120831	H	3.35	217.8	3.80
			Ks			3.91
2707	5480640	120831	H	3.77	181.5	3.55
			Ks			3.72
2707	5480640	120831	H	5.82	198.5	3.76
			Ks			3.84
2707	5480640	120831	H	8.21	350.9	5.22
2707	5480640	120831	H	8.67	224.5	2.34

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
			Ks			2.07
2707	5480640	120831	H	9.89	112.9	4.41
			Ks			4.80
2717	9467404	140819	Ks	8.71	332.7	6.50
2755	3545135	120807	H	6.81	194.7	4.86
			Ks			4.77
2771*	11456382	130603	H	3.71	309.5	5.98
2771*	11456382	130603	H	4.36	77.4	7.78
2771*	11456382	130603	H	8.01	133.9	8.63
2771*	11456382	130603	H	9.90	21.0	7.78
2795	5041569	140910	J	4.30	137.7	4.59
2795	5041569	140910	J	6.88	235.0	1.68
			Ks			1.63
2795	5041569	140910	J	7.52	25.7	2.44
			Ks			2.37
2795	5041569	140910	J	8.60	88.1	4.40
2795	5041569	140910	J	8.95	196.7	3.75
			Ks			4.30
2795	5041569	140910	J	9.81	194.2	2.99
			Ks			2.85

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
2835	5436338	130603	H	4.87	223.2	5.68
2835	5436338	130603	H	5.40	158.8	5.12
2838	6607357	120802	H	1.81	197.8	4.40
			Ks			4.91
2838	6607357	120802	H	5.16	342.0	7.60
2838	6607357	120802	H	7.23	155.7	5.37
2838	6607357	120802	H	7.98	134.3	2.79
			Ks			4.54
2838	6607357	120802	H	7.98	25.9	6.44
2838	6607357	120802	H	8.49	103.4	6.83
2838	6607357	120802	H	8.67	327.5	6.80
2838	6607357	120802	H	9.96	24.8	7.73
2857	6345732	120807	H	8.39	280.8	6.89
2991*	4848424	130605	H	4.01	90.5	6.14
2991*	4848424	130605	H	4.63	262.3	5.27
2991*	4848424	130605	H	6.83	222.1	4.61
2991*	4848424	130605	H	7.05	153.6	4.46
2991*	4848424	130605	H	7.90	252.2	6.19
3029	5903749	120801	H	0.24	263.2	0.39
3029	5903749	120801	H	1.75	355.6	5.72

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
3029	5903749	120801	H	2.56	3.8	3.95
			Ks			3.97
3029	5903749	120801	H	3.85	255.0	6.62
3029	5903749	120801	H	4.73	309.2	7.02
3029	5903749	120801	H	5.51	244.2	5.92
3029	5903749	120801	H	5.61	95.8	4.45
3029	5903749	120801	H	6.22	285.8	7.32
3029	5903749	120801	H	6.60	176.8	6.66
3029	5903749	120801	H	7.52	100.0	3.39
			Ks			3.62
3029	5903749	120801	H	8.60	326.8	5.58
3029	5903749	120801	H	8.74	59.1	3.93
			Ks			4.30
3029	5903749	120801	H	8.84	5.2	6.93
3029	5903749	120801	H	9.12	104.9	6.12
3029	5903749	120801	H	9.61	222.9	5.69
3029	5903749	120801	H	9.87	92.4	4.45
3246	9885417	140718	Ks	3.77	127.5	7.41
3818	6515722	140817	Ks	1.84	271.0	5.43
3818	6515722	140817	Ks	3.75	75.2	6.65

Table 4.1 (cont'd): Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter	Separation (arcsec)	Pos. Angle (deg)	Δm (mag)
3818	6515722	140817	Ks	9.77	159.7	2.24
3913	10281221	140708	J	9.95	250.4	5.24
4928	1873513	140718	Ks	8.82	303.1	4.78

Note. — * indicates a KOI that is no longer a planet candidate. Typical errors are 0.05″ for separation, 1 degree for position angle, and 0.1 mag for Δm.

Table 4.2: KOIs with No Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter
92	7941200	130602	H
110	9450647	140817	Ks
142	5446285	130602	Ks
144	4180280	140708	Ks
161	5084942	130602	H
174	10810838	130602	H
318	8156120	130603	H
323	9139084	130603	Ks
367	4815520	130429	H
388	3831053	140816	Ks
432	10858832	140817	Ks
470	9844088	140818	Ks
503	5340644	140816	Ks
526	9157634	140819	Ks
537	11073351	140718	Ks
580	8625925	140817	Ks
585	9279669	140819	Ks
660	6267535	140910	Ks
766	11403044	140718	Ks
783	12020329	140817	Ks

Table 4.2 (cont'd): KOIs with No Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter
844	6022556	140816	Ks
851	6392727	140816	Ks
916	8628973	140910	Ks
991	10154388	130604	H,Ks
1164*	10341831	120831	H
1303	10867062	120829	H
1430	11176127	130602	H
1647	11153121	120831	H,Ks
1688	6310636	120830	H
1747	7032421	120828	H
1786	3128793	120830	H
1813	9455325	140909	Ks
1815	9872283	140910	Ks
1820	8277797	140708	Ks
1832	11709244	141005	Ks
1833	11853878	141007	Ks
1839	5856571	141007	Ks
1867	8167996	120829,141005	H,Ks
1925	9955598	130429	H
1930	5511081	120806	H,Ks

Table 4.2 (cont'd): KOIs with No Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter
1960	6949061	130604	H
1962	5513648	130429,130604,140816	H.Ks
2005	6665512	130603	H
2035	9790806	141007	Ks
2037	9634821	120806	H,Ks
2046	10663396	120830	H
2069	11360571	130603	H
2148	6021193	120829	H,Ks
2175	9022166	120831	H
2233*	8963721	120831	H
2272	9654875	120829	H
2276	3458028	140819	Ks
2287	9718066	120830	H,Ks
2306	6666233	130604,141007	H,Ks
2352	8013439	120831	H
2593	8212002	130601	Ks
2668	5513012	141006	Ks
2675	5794570	120829	H,Ks
2687	7202957	120803	H,Ks
2694	9475552	140816	Ks

Table 4.2 (cont'd): KOIs with No Visual Companions Within 10″

KOI (#)	*Kepler* ID (#)	Date of observation (YYMMDD)	Filter
2875	12170648	141006	Ks
2950	6028860	140910	Ks
3097	7582689	120801	H
3848	12784167	140819	Ks
3913	10281221	140708	J,Ks
6188	8822421	140817	Ks
7016	8311864	140708	Ks

Note. — * indicates a KOI that is no longer a planet candidate.

4.6 Further Work

The unique combination of our single-planet target selection (§4.2) and the depth of our detections (Table 4.1) open many possibilities for future science, some of which are discussed in §6.3.2. Our underlying statistical framework, which is able to incorporate large error bars and non-detections, grounds our endeavor to observationally constrain the importance of dynamical evolution processes that produce single-planet systems through interactions with a bound companion. As such, we must carefully quantify our errors and upper limits. The first effort to this end will be to test the dependence of our Δm measurements on both the inclusion of individual image frames and the photometry pipeline that we have used; with this information, we will be able to compute error bars that include some of the systematic PSF modeling error. Detection lower limits will be provided by sensitivity curves calculated from the residual images that we have already constructed; relatedly we will compare the information content of our observations to those from other work by computing the BSC parameter introduced in Lillo-Box et al. (2014). Combining our color information with that from

the optical follow-up surveys will provide a preliminary companionship probability based on co-location in the three spatial dimensions and the density of background sources provided by galactic stellar population models. Finally, to assist the *Kepler* planet characterization effort, we will provide a correction to the planetary radii given the contamination from the detected nearby sources.

With the companionship probability constrained, we can then start to test for differences between the planet population around stars with and without likely bound companions. An example of this is provided below, where we have compared the period distribution of the innermost planet orbiting KOIs with or without visual companions at various distances. According to two of the most common non-parametric two-sample tests used to test for differences between observed population distributions (the Kolmogorov-Smirnov test and the Anderson-Darling test), none of the samples' distributions are different on a statistically significant level (all p-values > 0.5). However, it is not yet clear whether the period distribution of planets with bound large-distance perturbers is the same as that for planets without these perturbers; this statistical null result could be due to the fact that the effort described above has not yet been completed, and so there are non-binary stellar systems present in all of these samples. Combining the results from all the follow-up surveys listed in §4.1.1 can also augment the statistical power of this sample size and help test for differences between the planet radius, stellar radius, and other physically relevant distributions of the binary and non-binary host star samples. On the other hand, comparing the transiting planet properties with *Kepler*'s eclipsing binary distributions provides another test of the false positive

rate for these single-planet systems that is very complementary to the efforts outlined in §1.2.3. With these detailed analyses we can start to place quantitative constraints on the fraction of planetary systems which experienced secular dynamical evolution due to a distant, massive companion, and the probability that an individual planetary system could have undergone this process in its past.

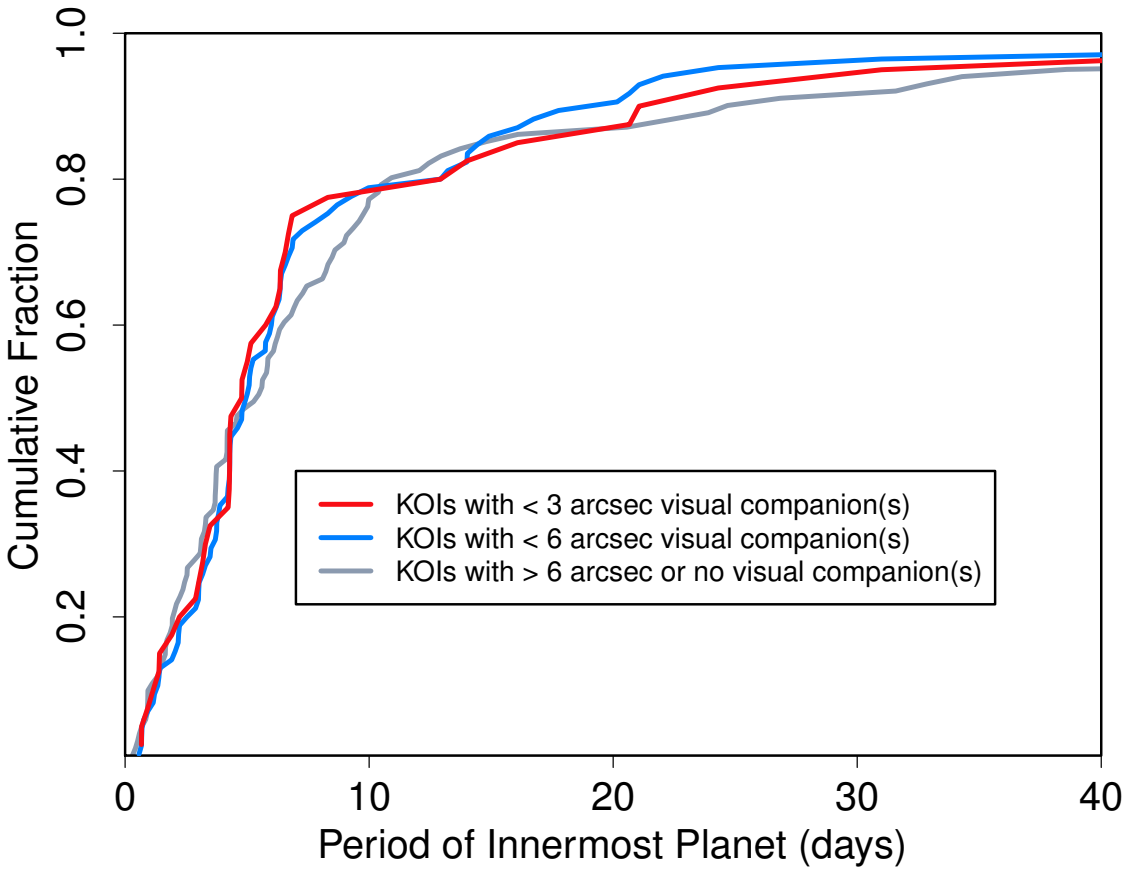

Figure 4.3: Cumulative period distributions of the planets orbiting KOIs with or without visual companions. According to both the Kolmogorov-Smirnov and the Anderson-Darling two-sample tests, the null hypothesis that the distributions come from the same population cannot be rejected: the $< 3''$ vs. $> 6''$ comparison gives p-values of 0.57 and 0.83 for the two different tests, respectively, and the $< 6''$ vs. $> 6''$ comparison gives p-values of 0.66 and 0.63.

Part II

Theory-Driven Characterization of *Kepler*'s Sub-Neptunes

Chapter 5

The Composition Distribution of *Kepler*'s Sub-Neptunes

5.1 Introduction

The *Kepler Mission* has found thousands of planetary candidates with sizes between that of Earth and Neptune (Rowe et al., 2015; Burke et al., 2014; Batalha et al., 2013; Borucki et al., 2011). Considering that no such planets exist in our own Solar System, this discovery elicits fundamental questions about their nature: are these planets scaled-up versions of Earth, scaled-down and irradiated versions of Neptune, or something in-between? What is the "typical" composition of this planet population, and what is the range of possibilities, as constrained by the planets we have observed? At what radius is the expected transition between rocky and gaseous compositions?

Addressing such population-wide inquiries about bulk compositions requires

two tools: first, models of internal structures which relate an individual planet's composition to its measured radius, and second, a statistical framework which combines information about individual members of a population into an inference about the whole, all while appropriately accounting for individual uncertainties. The former has been studied by a number of authors, as summarized below; the latter, however, has received limited treatment in the exoplanet literature. Here we provide an exoplanet-specific example of one such statistical tool commonly used for population studies in other fields, and in doing so answer questions about the range and distribution of compositions for these sub-Neptune-sized planetary candidates.

5.1.1 Modeling Sub-Neptune Interior Structures

Theoretical modeling of exoplanet interiors has a substantial history, starting with models that were developed to understand the structure and evolution of gas giants (e.g. Fortney et al., 2007; Marley et al., 2007). As recent surveys have uncovered ever smaller extrasolar planets, these models have been extended to the new population of sub-Neptune-sized planets. Studies of such low-mass planets include investigations of "ocean worlds" (Léger et al., 2004), low-density irradiated exo-Neptunes (Rogers et al., 2011), and scaling relations between mass and radius for sub-Jovian planets of varying compositions, including iron, silicates, water ice, carbon compounds, hydrogen/helium, and various combinations thereof (Seager et al., 2007; Fortney et al., 2007; Valencia et al., 2007a).

These models have been applied to numerous individual Neptunes and sub-Neptunes, such as GJ 876d (Valencia et al., 2007b), CoRoT-7b (e.g. Léger et al., 2009;

Valencia et al., 2010; Jackson et al., 2010), GJ1214b (e.g. Charbonneau et al., 2009; Rogers & Seager, 2010b; Nettelmann et al., 2011; Valencia et al., 2013), and the Kepler-11 system (Lopez et al., 2012). As the number of transiting planets with well-determined masses has grown, these models have been applied to ever-larger samples (e.g. Howe et al., 2014; Lopez & Fortney, 2014). However, these studies have fallen short of quantifying the relative numbers of planets at each composition in the underlying population, as such an effort requires correcting for *Kepler*'s survey biases and incompleteness and accounting for the uncertainties in the individual planet parameters. Fortunately, the statistical tool presented in §5.1.2 and described in detail in §5.3 enable us to take these crucial considerations into account, and to derive a quantitative composition distribution for the first time.

Inferring a planet's bulk composition from its mass and radius is of course a highly degenerate problem made worse by the possible choices for the number and type of layers in the planet's interior (Valencia et al., 2007a; Rogers & Seager, 2010a). Nevertheless, we can derive some guidance for how to best address this problem and make progress on answering these population-wide composition questions by noting a few salient characteristics of the overall low-mass planet population. First, a substantial fraction of these planets have radii that are just too large to be explained by rock/ice/water combinations (e.g, Lopez & Fortney, 2014; Rogers, 2015). Second, mass constraints for a few dozen sub-Neptunes indicate that planets at the same radii can vary in mass by a factor of $\sim 2-4$ (Marcy et al., 2014; Weiss & Marcy, 2014), hinting at significant compositional variability within this population. Finally, conclusively rock-

like bulk density constraints have been obtained for a few small planets, most notably Kepler-36b (Carter et al., 2012) and the highly irradiated Kepler-78b (Pepe et al., 2013; Howard et al., 2013) and Kepler-93b (Dressing et al., 2014).

Based on these observations, we adopt a few key assumptions which allow us to move forward with our work to infer statistical (versus individual) compositions for *Kepler*'s sub-Neptune population. First, the existence of measured masses and radii that can be fit with decisively rocky compositions while others require non-negligible amounts of hydrogen and helium motivates us to assume that all of these low-mass planets are a part of the rocky/H+He continuum. By framing the problem in this way, we are effectively considering the end-member case where all planets contain some amount of hydrogen gas over a rocky core. Therefore, we approach these planets as gaseous until their envelope mass fractions are so low that they can be called rocky[1].

Of course, this does not mean that "water worlds", i.e. planets with either a distinct water layer or with water vapor comprising a substantial percentage of the gaseous envelope, could not exist. Indeed, if photoevaporation plays a significant role in shaping this irradiated planet population, planets which are shown to exclusively lie in the radius-flux "occurrence valley" predicted by Lopez & Fortney (2013) and Owen & Wu (2013) are likely such water worlds. Nevertheless, with core accretion as a reasonable proposal for the formation of these sub-Neptunes and with protoplanetary disks composed primarily of hydrogen, the most straightforward explanation for the substantial compositional variation implied by measured masses and radii is variation in the

[1]From an astronomer's perspective, this is determined by the typical error bars on the planets' radii in the best-case scenario where the stars are well characterized, which translates to a gaseous mass fraction of $\lesssim 0.1\%$

accretion and loss of hydrogen (given that water atmospheres have smaller scale heights than H+He, a somewhat extreme dynamic range in the processes of rock/ice coagulation, disk migration, and water differentiation and evaporation is needed to produce an entire population of water worlds which match the full range of these observations). Before venturing too far with philosophical invocations of Occam's Razor, however, we acknowledge that a sub-population of planets with water-dominated atmospheres could very well exist. We will investigate this idea in future work, but first need to lay our groundwork for a statistical treatment of planet compositions using the simpler scenario that we assume here.

With the postulate that gaseous envelopes tend to dominate the non-rocky portion of the planet's structure, we can adopt a two-component interior structure and, for now, set aside the problem of compositional degeneracy that arises from models with three or more layers. Even so, a large amount of theoretical uncertainty in the intrinsic luminosity of these planets remain. Fortunately, coupling interior structure models to atmospheric radiative transfer models (e.g. Fortney et al., 2007; Guillot, 2010) enables tracking of the thermal cooling of a planet's interior as it ages, eliminating the need to marginalize over the internal energy (Lopez et al., 2012; Lopez & Fortney, 2014). When applied to highly irradiated sub-Neptunes as done in Lopez & Fortney (2014), these thermally evolving models result in radii that are more sensitive to the fraction of a planet's mass that is in a hydrogen and helium envelope than to the total mass. This has significant implications, as mass measurements are not needed to get a sense for this composition parameter, and the information content in the *Kepler* radius distribution

can be maximally leveraged for such studies.

5.1.2 Statistical Treatment of Planet Populations

The population characterization studies outlined in §1.3 use a range of statistical techniques, including the intuitive yet idealized inverse efficiency method (Howard et al., 2012; Petigura et al., 2013); linear regression on binned, mean estimates of very uncertain, intrinsically dispersed individual points (Weiss & Marcy, 2014); Monte Carlo approaches (Wolfgang & Laughlin, 2012; Fressin et al., 2013); maximum likelihood that incorporates survey incompleteness (Youdin, 2011; Tremaine & Dong, 2012; Dong & Zhu, 2013); and non-parametric kernel density estimation (Morton & Swift, 2014). Each of these studies treat error in the observed quantities of individual planets differently, but none incorporate them in a way that produces rigorous posterior estimates of the population parameters of interest. Given that the goal of such population studies is to characterize the population given the observed data, the quality of this data should play a large role in the inference of the population parameters.

Hierarchical Bayesian modeling (HBM) is very naturally suited to this problem, and has been in use for decades by many fields, including bioinformatics and political science, whose key science questions involve inferring characteristics of a population from noisy observations of individual members. The HBM framework is very general, and its usefulness extends to a number of commonly encountered problems in astronomy; we refer the reader to Loredo (2007) and Loredo (2013) for a discussion of multi-level modeling in a general astronomical context, and to §5.3 for an overview of its capabilities for the science goals of this work.

The promise that HBM holds for exoplanet population studies has only recently been realized. The first instance of HBM in the exoplanet literature is Hogg et al. (2010); they derive an importance sampling algorithm which incorporates posterior samples that have already been computed for individual planets into inferences about the population, and they apply this algorithm to the planetary eccentricity distribution. Foreman-Mackey et al. (2014) and Rogers (2015) use this algorithm to infer the occurrence rate of planets as a function of period and radius, and to infer the radius at which super-Earths transition from gaseous to rocky compositions, respectively. This work, on the other hand, is the first study to perform full hierarchical Bayesian modeling where simultaneous inferences on both the population and the individuals are made; the details are laid out in §5.3.4 - 5.4.2.

In this chapter we present the first quantitative distribution of sub-Neptune compositions, which we define as the fraction of a planet's mass that exists in a hydrogen and helium envelope around an Earth-like rocky core that can vary in mass. In §5.2, we describe how the *Kepler* planet candidates sample used for this work was selected. In §5.3 we explain what hierarchical Bayesian modeling is and detail the specifics of the model that we use to obtain the sub-Neptune composition distribution presented in §5.4. We discuss the implications of these results in §5.5, and conclude in §5.6.

5.2 Using the *Kepler* Objects of Interest

The composition distribution that is the subject of this work is solely constrained by the *Kepler* planet candidates and their radius distribution; no additional

follow-up observations, such as radial velocity measurements, were used to obtain these results. As discussed in §5.1.1, this simplification is enabled by the theoretical discovery of Lopez & Fortney (2014), which found that when the thermal evolution of a rock-dominated body is coupled with an internal structure model of its gaseous envelope, its radius is primarily determined by the mass fraction of that envelope (we denote this quantity as f_{env}). The construction of the *Kepler* Objects of Interest (KOIs) catalog that we use here is detailed in §1.2.

5.2.1 Selecting a Complete Subsample

As described in §1.2.2, all statistical surveys, and especially those which endeavor to obtain quantitative comparisons with theory, must account for the survey reliability and completeness. We detail those efforts here. To select our sample, we begin with the cumulative *Kepler* Objects of Interest (KOI) table available at the NASA Exoplanet Archive (Akeson et al., 2013), which at the time of access (December 2, 2013) consisted of the Q1-12 catalog (Rowe et al., 2015), a heterogenous list of KOIs identified in the first 12 quarters or less of *Kepler* data. The heterogeneity arises from the fact that, in general, the higher signal-to-noise (S/N) events are identified with fewer data and a less mature vetting process (overviewed in §1.2). Furthermore, the reported planet parameters in the cumulative catalog are derived from different total amounts of data.

Even though this list does not yet represent the uniform sample that is ideal for statistical studies, it is the best-knowledge catalog to date. Uniformity in planet parameters, if not in the planet candidate (PC) disposition itself, can be improved by

matching the KOIs to the latest Threshold Crossing Events (TCEs) (Tenenbaum et al. 2014; see §1.2 for a discussion of the differences between these lists; vetting for the Q1-16 planet candidate catalog had not yet started at the time of this analysis). Matching Q1-12 PCs to Q1-16 TCEs also ensures that the planet parameters used in our analysis are those derived with the best-knowledge stellar parameters: the Q1-16 stellar properties are described in Huber et al. (2014), hereafter referred to as Hub14.

Starting with the 3601 KOIs listed as PCs in the Q1-12 catalog, we retain 3322 PCs whose stars have a Q1-16 TCE within 1% of the PC period P and an epoch modulo P within $0.05 * P$ days of the PC epoch. Half of the discarded PCs do not have any Q1-16 TCEs identified for that target star. This could be due to the pipeline's mistaken removal of short-period, high-S/N transits via the narrow-band oscillation filter described in Tenenbaum et al. (2014), or to strong transit timing variations (TTVs), or to PCs that are actually false alarm detections. The other half of the discarded PCs with non-matching periods and epochs can also be explained by TTVs or false alarms, or by the more common circumstance where the pipeline identifies a harmonic or sub-harmonic of the true transit signal (Tenenbaum et al., 2014).

In this work we characterize the compositions of sub-Neptune planets, so we limit ourselves to the 2572 PCs with 1 $R_\oplus < R_{pl} < 4$ $R_\oplus$. Because we are using the Q1-16 TCE parameters, these radii are derived from the Q1-16 data using the Q1-16 stellar parameters of Hub14. The remainder of our sample cuts arise from concerns about the completeness of this sub-Neptune sample. Because our analysis method (overviewed in §5.3, detailed in §5.3.4) automatically folds the shape of the PC radius distribution into

our result on the composition distribution of sub-Neptunes, we must take precautions to ensure that this PC radius distribution is as close to the true planet radius distribution as possible. Figure 5.1 shows the radius distribution of our final sample compared with the total distribution of Q1-12 PCs. Per the discussion in §1.2.2, we expect that the full sample is less complete at smaller radii; the black points, which denote the effective completeness correction made by choosing a complete subsample, illustrate that this is indeed the case. Thus, the cuts described below are effective in minimizing the detection biases present in the larger catalog.

To define a host star sample around which the detection of sub-Neptune planets should be complete, we use the standard S/N calculation for a transiting planet (see, for example, Wolfgang & Laughlin 2012) and *Kepler*'s detection criteria of 7.1σ. We find that a $R_\star < 1.2$ $R_\odot$ star with noise < 100 ppm on transit duration timescales that had been observed continuously for three years should be complete for planets with $P < 25$ days and $R_{pl} > 1.2$ $R_\oplus$. This radius cut encompasses the vast majority of planets which are conservatively expected to still have a gaseous envelope, and so preserves completeness of the planets which contribute to our composition distribution. We therefore restrict our sample to main-sequence host stars ($\log(g) > 4.0$) with $R_\star <$ 1.2 $R_\odot$ and a CDPP value that when scaled to the duration of the planet's transit is less than 100 ppm. We further require that the host star have been observed for all 12 quarters. With the final cut on period, we retain a sample size of 215 sub-Neptune sized planets within ~ 0.15 AU of their host stars (Figure 5.1).

Although this careful selection of the host star sample accounts for detection

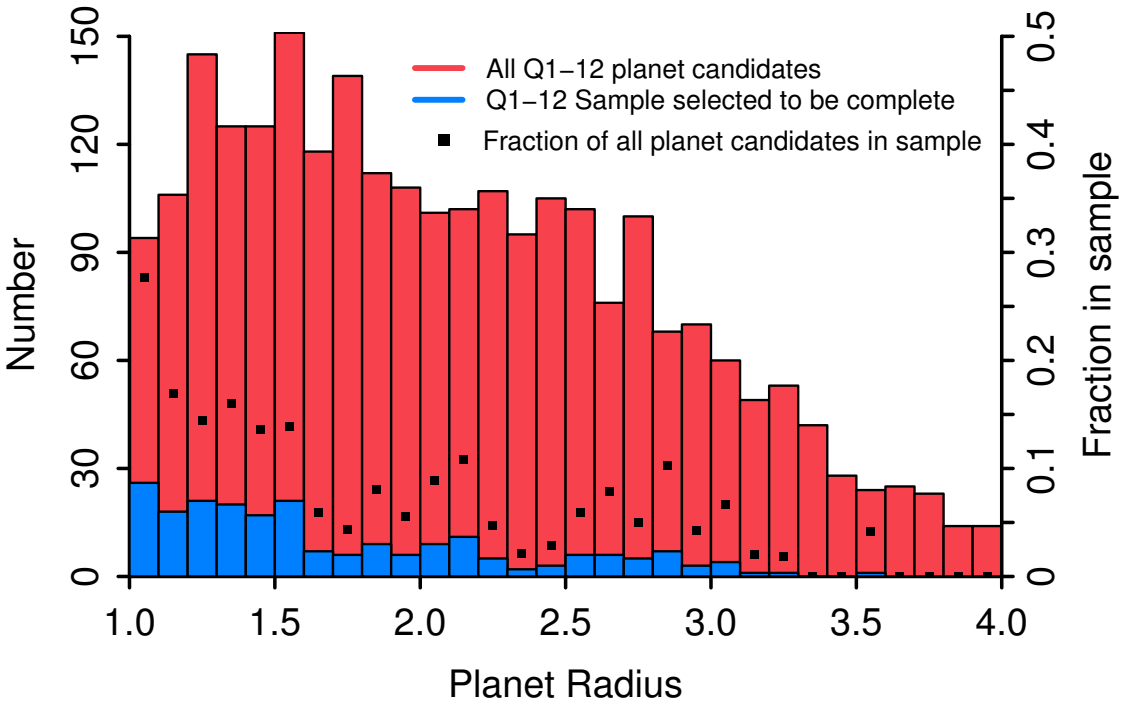

Figure 5.1: Radius distribution of our subsample (blue; $N = 215$) of sub-Neptune planet candidates (PCs) compared with the total distribution of Q1-12 PCs in this size range (red; $N = 2572$). Detection biases cause fractionally fewer small planets to be found, especially at longer periods. Carefully restricting the parent star sample and imposing a period cut, as was done to create the subsample, can mitigate these biases (see §5.2.1 for details). The black points corresponding to the right y-axis quantify this mitigation, showing the fraction of PCs in that radius bin from the total Q1-12 catalog which made it into our more complete subsample.

bias, pipeline incompleteness is still a concern. To assess how much of an effect this could have on our results, we scale the Q1-16 transit model SNRs of our Q1-12 sample to the time baseline over which they were detected, and display the results in Figure 5.2. We note that only 15% of our sample has SNR < 15, where pipeline incompleteness becomes significant; of these, 95% have $R_{pl} < 1.6$ $R_\oplus$, which we show in §5.5.3 are most probably rocky given our "best-fit" composition distribution. Therefore, pipeline

incompleteness does not affect the distribution of gaseous mass fractions that we infer from *Kepler*'s irradiated sub-Neptune population (§5.4.1).

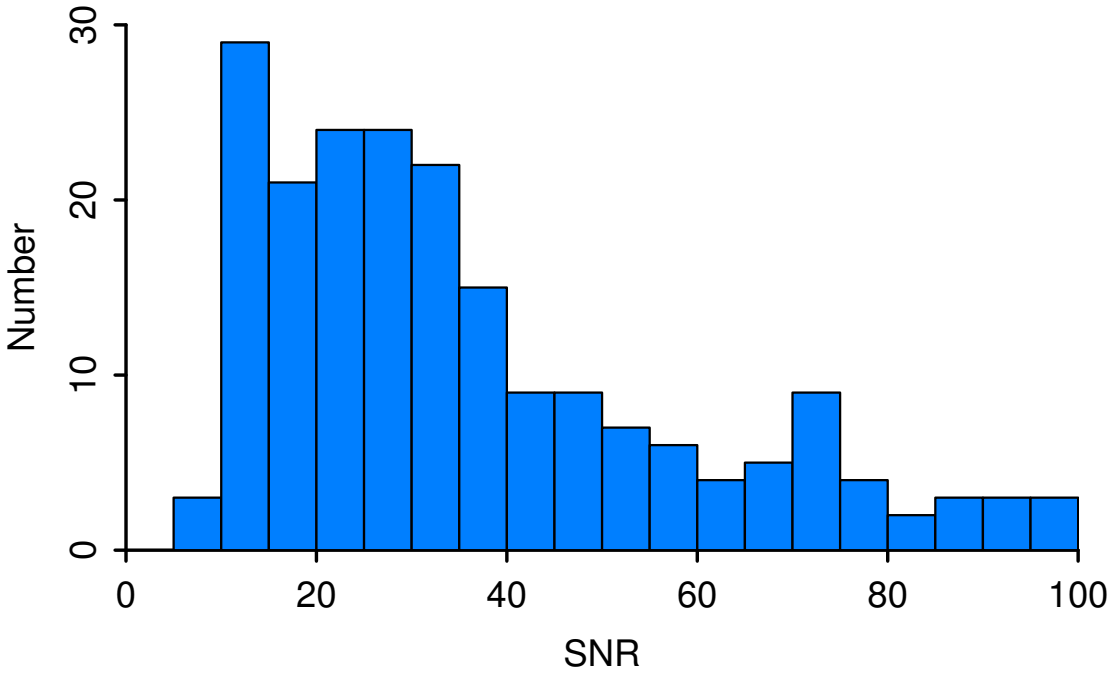

Figure 5.2: Distribution of the Q1-16 signal-to-noise ratios of the majority of our sample, scaled to the Q1-12 detection baseline (14 PCs have scaled SNR > 100 and are not displayed). 32 PCs have SNR < 15, where pipeline incompleteness becomes significant; of these, two have $R_{pl} > 1.6$ R$_\oplus$. Because only $\sim 1\%$ of our sample suffers from pipeline incompleteness while having nonzero gaseous envelope mass fractions, our composition distribution (§5.4.1) is not affected by this otherwise problematic sample bias.

5.3 Methods: Characterizing Planet Compositions via Statistical Modeling

The goal of this work is to understand the range of gaseous envelope mass fractions that *Kepler*'s super-Earths and sub-Neptunes can possess. In this section,

we motivate why hierarchical Bayesian modeling (HBM) is such a natural approach to this problem, discuss some of its advantages over other methods, and detail the specific model that we use to infer the sub-Neptune composition distribution. Due to the limited use of HBM in the exoplanet literature we spend significant time explaining the reasons, context, and application of this choice, but for the hurried reader we provide the following summary:

- Hierarchical Bayesian modeling is the natural choice for constraining the population distributions of exoplanet properties (such as compositions or radii), when those properties are either unobserved (compositions) or possess significant errors (radii).

- HBM is also the natural choice when the priors on individual exoplanet properties are expected to have an intrinsic scatter instead of one true value, where the scatter is due to some physical variation among the population and is of scientific interest.

- HBM provides posteriors on both the population parameters (e.g., the mean of the composition distribution) and on the individual parameters (e.g. an individual planet's composition), thereby enabling simultaneous inference on individual planets and the population as a whole.

- By relating individuals to each other through this hierarchical framework, HBM provides posterior estimates of individual exoplanet properties which have smaller variance than if multiple individual Bayesian analyses were performed indepen-

dently. This is called "shrinkage" and is illustrated in §5.4.3.

- HBM is a straightforward extension of regular Bayesian modeling, requiring only the definition of conditional probability and a slight shift in interpretation, and so uses the same basic computational algorithms such as Markov Chain Monte Carlo.

- As with all Bayesian analysis, HBM enables prediction of future observations by marginalizing the likelihood over the posterior distributions. We present the sub-Neptune posterior predictive composition distribution in Figure 5.5.

The discussion below details the application of hierarchical, or multi-level, modeling to exoplanet compositions; for a more general discussion of the past use and future promise of multi-level modeling in astronomy, we refer the reader to Loredo (2007, 2013).

5.3.1 Choosing an Appropriate Statistical Framework

To understand why we have chosen HBM to solve this problem, we must first understand how these planets' compositions relate to the quantities that *Kepler* measures. Most simply, a sub-Neptune's gaseous mass fraction sets its radius, as Lop14 showed that these planets' compositions dominate over other factors in determining their size; the radius, in turn, is primarily derived from the depth of the transit signal, which is the quantity that *Kepler* directly observes. In practice, however, several other quantities become important to include in order to accurately infer planetary compositions from their transit parameters; the relationships between them for a single planet candidate

are shown in Figure 5.3.

The hierarchical structure of this problem is immediately apparent. Having such a multi-tiered relationship between relevant quantities does not necessarily require a hierarchical Bayesian framework, however. Simple inversion of the problem and standard error analysis is sufficient if the relationships are deterministic (that is, they can be summarized as a function that maps one set of input values onto one output value) and if the values of the quantities themselves are well known with errors that are either small or well-behaved (i.e. symmetric and uncorrelated).

For the problem outlined in Figure 5.3, the relationships could indeed be deterministic (but see the discussion about likelihoods below and the full problem outlined in Figure 5.4 and Equations (5.8) - (5.9)). On the other hand, the values of many of the quantities in Figure 5.3 are neither well known nor are their uncertainties well-behaved. For example, the *Kepler* pipeline described in §1.2 produces biased and poorly constrained estimates for impact parameter b (Rowe, private communication). Even more problematic is the dependence of these quantities on stellar parameters: Hub14 illustrates that the current state of the observations of *Kepler*'s target stars leads to large, asymmetric uncertainties on $R_\star$ and $M_\star$, and can even make them multimodal (see Figure 8 of that paper). As a result, the envelope fractions f_{env} cannot be straightforwardly calculated from the observed transit depths δ under the assumption of small errors, and a more sophisticated analysis is warranted.

To incorporate our observational uncertainty, we must relax the requirement of deterministic relationships and allow probabilistic, or stochastic, relationships within

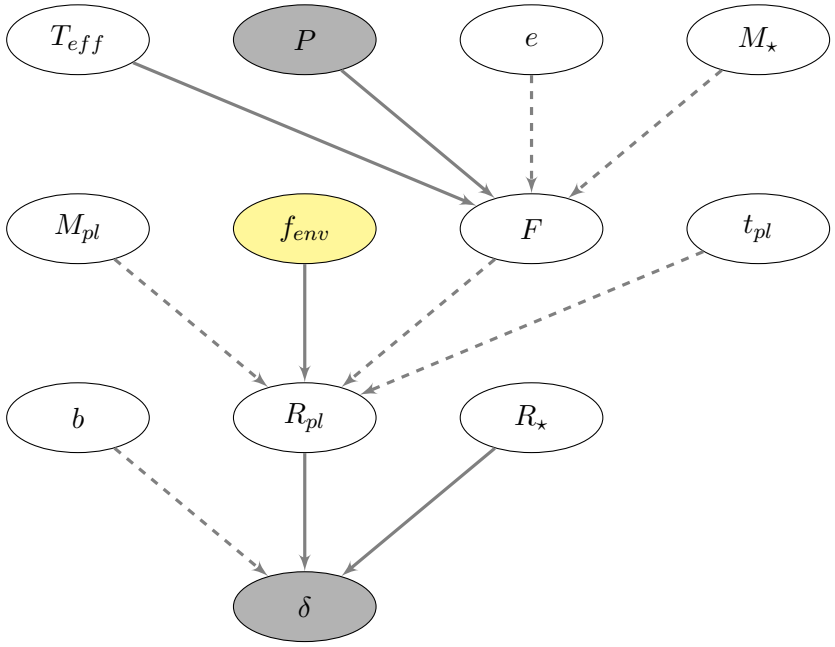

Figure 5.3: The relationships between the quantities that *Kepler* observes (gray ellipses) and the quantity of interest in this work (highlighted in yellow). This diagram represents the flow of information for a single planet candidate. First-order relationships, i.e. those that dominate the value of the resulting quantity, are denoted as solid lines whereas second-order relationships are represented by dashed lines. Note that the mapping from the second line to the planet radius is given by the models of Lop14. These quantities are defined as follows:

T_{eff} = effective temperature of host star
P = period
e = eccentricity
$M_\star$ = mass of host star
M_{pl} = total mass of planet
f_{env} = fraction of M_{pl} existing in a gaseous H+He envelope
F = stellar flux incident on the planet
t_{pl} = age of planet
b = impact parameter
R_{pl} = radius of planet
$R_\star$ = radius of host star
δ = transit depth

the structure of Figure 5.3. This is accomplished by computing the likelihood function,

which describes how probable the data are under their measurement uncertainty, given

different values for the model parameters. Choosing the appropriate likelihood function requires knowledge about how the measurement errors behave; often it is assumed that they follow a Gaussian distribution, meaning that the measured values are normally distributed around the true value. Answering the question of interest, i.e. "what is the gaseous envelope mass fraction of planet X", then involves inference, where one identifies the parameter values which "best fit" the data.

"Best fit" parameter values can be found by maximizing the likelihood function directly, which gives an estimate of the "true" value exhibited by nature. Alternatively, by shifting one's interpretation of the likelihood to allow for uncertainty in the true parameter values, one can combine the likelihood with some prior information to create a posterior distribution of likely parameter values. In practice, the former method of maximum likelihood (ML) often manifests as calculating χ^2, which requires the aforementioned assumption of normally distributed errors (note, however, that ML can be performed for any arbitrary likelihood function, by solving for the parameter values at which such a likelihood is maximized). In contrast, the latter method of Bayesian inference usually involves Markov Chain Monte Carlo (MCMC) simulations, wherein a sequence of posterior probabilities is numerically computed in a way that optimally explores the range of parameter values allowed by the data.

While the choice between using ML and Bayesian inference is often a matter of philosophical preference (see Loredo 2013 for an in-depth discussion on the philosophical differences between frequentist and Bayesian approaches), ML has the strongest computational advantage over Bayesian methods when one has no prior information and

when the likelihoods are easy to write down, analytically tractable, and do not involve too many parameters. In that case, the matrix inversion required to find the best-fit values can be performed quickly and efficiently, and the confidence intervals in those best-fit values can be analytically computed.

Unfortunately, such an analytic treatment is not possible for the problem we endeavor to solve in Figure 5.3, given the necessarily numerical calculation of the stellar parameters and their errors, which are non-Gaussian. In such a situation, ML would involve computing likelihoods on a grid of parameter values, and error bars would be interpreted as the range of parameter values which enclose the maximum likelihood estimate for 68% of the datasets. Not only is the latter task difficult to do with a single dataset, but this approach is much less computationally efficient than a Bayesian treatment involving MCMC, as the Markov Chain spends less time exploring parameter space that has low probability of matching the data. This computational consideration, in combination with the realization that a Bayesian approach is better suited to our problem, where we only have a single list of planet detections from *Kepler* and significant uncertainty about the true physical parameters of the planet population, guides our choice of a Bayesian framework for this study. In doing so, we also enable the incorporation of prior information, which can naturally be extended into the hierarchical structure appropriate for this problem, as explained in §5.3.3.

5.3.2 Applying Bayes' Theorem

The basic Bayesian framework is readily summarized with the following interpretation of Bayes' Theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})}, \tag{5.1}$$

where $\boldsymbol{\theta}$ is the set of parameters that define the model (i.e. the quantities circumscribed by ellipses in Figure 5.3), $\boldsymbol{X}$ is the set of data values (i.e. the quantities circumscribed by rectangles in Figure 5.3), and $p(x|y)$ denotes the probability distribution function of quantity x at a given value of quantity y (in other words, the probability of x conditional on y). Inference occurs via the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$, which yields the probability of various parameter values given the data; the "best fit" values can be the mode, median, or some other central statistic of this distribution. Computing the posterior requires specifying the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ as described above and the prior distribution $p(\boldsymbol{\theta})$, which reflects previous information about how intrinsically likely different parameter values are; the normalizing constant $p(\boldsymbol{X})$ can be ignored when one uses MCMC to compute the posterior numerically, as the core of the MCMC algorithm involves computing posterior probability ratios within which this constant cancels.

To apply this framework to our problem, we note that transit depth δ is our primary observable quantity, so we set $\boldsymbol{X} = \{\delta\}$. $\boldsymbol{\theta}$ therefore denotes the rest of the unobserved quantities in Figure 5.3. A full Bayesian treatment would require specifying the joint prior probability function $p(R_{pl}, R_\star, b, M_{pl}, f_{env}, F, e, T_{eff}, M_\star, t_{pl})$ along with our likelihood $p(\delta|R_{pl}, R_\star, b, M_{pl}, f_{env}, F, e, T_{eff}, M_\star, t_{pl})$, but in practice the varying levels of importance in the relationships between the different quantities, as denoted by

the dashed vs. solid lines in Figure 5.3, allow us to simplify the problem. Accordingly, we hold constant the values of parameters that are related to a second-order quantity, thereby setting P to the observed value, $e = 0$,[2] and T_{eff} and $M_\star$ to the best-fit values determined by Hub14. We also set to fiducial values second-order parameters such as $b(= 0)$ and $t_{pl}(= 5\text{Gyr})$ that are not well constrained by the data.

With these modifications, our prior probability has simplified to $p(R_{pl}, R_\star, M_{pl}, f_{env}, F)$, but still has a nontrivial functional form given the hierarchical dependence between the parameters. Fortunately, we can put this intrinsic structure to use: rather than specify one distribution containing all of these parameters, we instead use the definition of conditional probability to derive a joint prior probability distribution in terms of conditional distributions. This definition states that

$$p(x, y) = p(x|y)p(y), \tag{5.2}$$

where $p(x, y)$ is the joint probability distribution of x and y (i.e. it states the probability of both of those x and y values occurring), $p(x|y)$ is the conditional probability distribution of x given y (i.e. at a set value of y, it states the probability of an x value), and $p(y) = \int p(x, y)dx$ is the marginal probability distribution of y (i.e. it states the probability of the y value occurring under all conditions). Therefore, we can split our joint prior probability distribution into a series of conditional and marginal distributions as appropriate given the structure of our problem:

[2]For these small planets which don't have measurable occultations and which yield relatively low SNR transits, individual eccentricities are not well constrained by the transit data alone.

$$
\begin{aligned}
&p(R_{pl}, R_\star, M_{pl}, f_{env}, F) \\
&\quad = p(R_{pl}|R_\star, M_{pl}, f_{env}, F) \\
&\qquad \times p(R_\star|M_{pl}, f_{env}, F)p(M_{pl}, f_{env}, F) \\
&\quad = p(R_{pl}|M_{pl}, f_{env}, F)p(R_\star)p(M_{pl})p(f_{env})p(F).
\end{aligned} \tag{5.3}
$$

Note that in simplifying the right-hand side we have assumed that M_{pl}, f_{env}, and F are independent of each other, that $R_\star$ is independent of M_{pl}, f_{env}, and F, and that the true, intrinsic radius of the planet R_{pl} is independent of $R_\star$; this is also reflected in the structure of Figure 5.3. The usefulness of this framework is that such dependencies can be effortlessly included in subsequent analysis should there be good reason to expect that they exist or are important.

To make any further progress, we must specify what functional forms these prior probabilities should take. $p(R_\star)$ is the distribution of allowed radius values for the host star, and is equivalent to the likelihood numerically calculated by Hub14 after it has been marginalized over all other stellar parameters; inclusion of this distribution among our prior information is how we are able to account for uncertainties in the stellar parameters. $p(R_{pl}|M_{pl}, f_{env}, F)$ represents the Lop14 sub-Neptune internal structure models. Because these models map a planet's mass, envelope fraction, and incident flux to a single radius value, this probability distribution is actually a delta function; this

is how we allow for deterministic relationships in our probabilistic model. Due to the simplifying choices we made above, we have also forced $p(F)$ to be a delta function. Implicit marginalization over these last two parameters with delta function probability distributions then allows us to write:

$$p(R_{pl}, R_\star, M_{pl}, f_{env}, F) = p(R_\star)p(M_{pl})p(f_{env}), \tag{5.4}$$

where R_{pl} will show up in the likelihood as a deterministic function of M_{pl}, f_{env}, and F.

This leaves specifying $p(M_{pl})$ and $p(f_{env})$. First we address planet mass: while the result of Lop14 — that sub-Neptune radii are relatively insensitive to their masses compared to the effect of the gaseous envelope mass fractions — is what inspired this work, considering the planets' mass is still important for the smallest envelope fractions and thus for the posited transition between gaseous and rocky planets. It is therefore necessary to retain consideration of the planet masses for this study. Unfortunately, we do not have mass measurements for every individual planet in our complete subsample of *Kepler*'s small planet candidates, and so we cannot specify a per-planet probability distribution for M_{pl} as we did for $R_\star$. However, we do have an idea of the mass distribution of the low-mass planet *population* from radial velocity surveys. Therefore, we can base our individual planet mass prior on the population distribution of masses; if we follow the RV surveys and choose a power law that is parameterized with the index

α, then:

$$p(M_{pl}) = \int p(M_{pl}|\alpha)p(\alpha)d\alpha$$
$$= C \int M^{\alpha} p(\alpha)d\alpha. \tag{5.5}$$

Parameterizing the prior of an individual quantity based on the distribution within the population is exactly what makes this particular Bayesian formalism hierarchical, and is why we have turned to HBM to solve this problem.

5.3.3 Hierarchical Bayesian Modeling

Mathematically, the general framework of HBM is a deceptively simple adjustment to Equation 5.1:

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\beta})p(\boldsymbol{\theta}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\boldsymbol{X})}, \tag{5.6}$$

where the difference between the set of individual parameters $\boldsymbol{\theta}$ and the set of population parameters $\boldsymbol{\beta}$, referred to as "hyperparameters", has been made explicit. Nevertheless, this rewrite, which is based only on the definition of conditional probability, makes a substantial difference in the interpretation of the problem, as one can now group individuals into populations, which both facilitates the characterization of the individual

and allows the individual data to inform the population hyperparameters. HBM thereby allows simultaneous inferences on both the parameters of the individual and of the population.

That said, HBM is not always necessary to answer the question that has been posed. In particular, the hyperparameters may not always be of interest, in which case they can be treated as "nuisance parameters" and marginalized over, as Equation 5.5 implies. Therefore, many hierarchical structures such as that in Figure 5.3 do not necessarily need an HBM treatment. The aspect of our problem which does require HBM is the specific question we have asked regarding compositions: because we want to infer the population distribution of compositions, we are interested in the analogous hyperparameters for f_{env}, and need the posterior to contain their distribution. Only HBM can provide such a posterior that incorporates both the parameters of the individual planets and the population hyperparameters.

Applying this general framework specifically to f_{env} means that we replace $p(f_{env})$ with $p(f_{env}|\mu, \sigma)p(\mu, \sigma)$, where μ and σ are the hyperparameters characterizing the composition distribution. The combination of computational convenience, the need for a distribution that can span several orders of magnitude, and the intuition that there should be fewer sub-Neptune planets with high envelope fractions leads us to choose a lognormal distribution for f_{env} (see §5.5.1 for an in-depth discussion of this choice), so that μ and σ are the mean and standard deviation of the population of $\log(f_{env})$ values.

This distribution does not factor in the expectation that significantly irradiated planets should have lost their envelopes, which we expect given the well-constrained

rocky compositions of Corot-7b (Jackson et al., 2010; Valencia et al., 2010), Kepler-10b (Batalha et al., 2011; Kurokawa & Kaltenegger, 2013), and Kepler-78b (Pepe et al., 2013; Howard et al., 2013). Ignoring the physics of evaporation could therefore lead to an unrealistic composition distribution for these close-in planets. That said, the photoevaporation of sub-Neptunes is an active area of theoretical research (e.g. Owen & Jackson 2012; Lopez et al. 2012; Lammer et al. 2013), and so we err on the side of a simple yet theoretically motivated prescription to arrive at a realistic result. In particular, we implement the mass loss threshold of Lopez & Fortney (2013), which is a scaling law for the incident flux a planet would need to have received from its host star to have lost half of its initial H+He envelope after several Gyr (F_{thresh}); it is based on the assumption of energy-limited hydrodynamic escape and depends primarily on the mass of the planet's core, to a power that varies slightly depending on the planet's composition. We model this irradiated sub-Neptune population by assigning a rocky composition ($f_{env} = 0$) to a planet if

$$F > F_{thresh} = (M_{core})^{\gamma}, \tag{5.7}$$

where F is the incident flux on the planet from its host star. Otherwise, the planet has a non-zero f_{env} which contributes to constraining the composition distribution. Given the theoretical uncertainties in this treatment of photoevaporation, we also allow γ to vary, thereby adding a fourth hyperparameter to our model.

5.3.4 Our Hierarchical Model

When written in the context of Bayes' Theorem, our full statistical model is the following:

$$
\begin{aligned}
& p(\boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{X}) \\
& \quad \propto \prod_{i=1}^{N} \Big\{ p(\delta_i | \sigma_{\delta,i}, R_{pl,i}, R_{\star,i}, M_{core,i}, f_{env,i}, F_i, \alpha, \mu, \sigma, \gamma) \Big\} \\
& \qquad \times \prod_{i=1}^{N} \Big\{ p(R_{\star,i}) p(M_{pl,i} | \alpha) p(f_{env,i} | \mu, \sigma) \Big\} \\
& \qquad \times p(\alpha) p(\mu) p(\sigma) p(\gamma), \qquad (5.8)
\end{aligned}
$$

where $\boldsymbol{X} = \{\delta_i, \sigma_{\delta,i}\}, \boldsymbol{\theta} = \{R_{pl,i}, R_{\star,i}, M_{core,i}, f_{env,i}, F_i\}$ and $\boldsymbol{\beta} = \{\alpha, \mu, \sigma, \gamma\}$ are defined in the caption of Figure 5.4 and in the above text, and we are now considering all of the planet candidates, with sample size N=215. Note that the normalizing constant $p(\boldsymbol{X})$ has not been written down, necessitating the expression of proportionality, and that we have assumed that all of the hyperparameters are independent of each other. To illuminate how this follows Bayes' Theorem, we point out that the first line of this equation is the posterior, the second line is the likelihood, the third line contains the prior distributions for the individual parameters, and the fourth line contains the priors on the hyperparameters.

Needless to say, this equation is unwieldy, as is the case with most hierarchical models. Graphical representations are therefore more often used to succinctly

communicate the problem; the graphical model corresponding to Equation 5.8 is shown in Figure 5.4. However, neither equation nor figure contain the details of the various probability distributions, and so we introduce another, more informative way of writing down our hierarchical model. In what follows, the quantities on the left-hand side are sampled from the distributions on the right-hand side; in other words, "$q \sim$" is shorthand for "$p(q) =$" where $p(q)$ is the probability distribution of the quantity q. The parameters which directly specify each probability distribution are located after the "|":

$$
\begin{aligned}
\delta_i | \sigma_{\delta,i}, \boldsymbol{\theta}, \boldsymbol{\beta} &\sim \text{Normal}\Big(\delta_i \Big| (R_{pl,i}/R_{\star,i})^2, \sigma^2_{\delta,i}\Big) \\
R_{pl,i} | M_{core,i}, f_{env,i}, F_i, \boldsymbol{\beta} &= g(M_{core,i}, f_{env,i}, F_i, \gamma) \\
R_{\star,i} &\sim \text{Gamma}\Big(R_{\star,i} \Big| a_i, b_i\Big) \\
f_{env,i} | \mu, \sigma &\sim \text{LogNormal}\Big(f_{env,i} \Big| \mu, \sigma\Big) \\
M_{core,i} | \alpha &\sim \text{Pareto}\Big(M_{core,i} \Big| -(\alpha+1), 0.5\Big) \\
\mu &\sim \text{Uniform}(-3.5, -1) \\
log(\sigma^2) &\sim \text{Uniform}(-4, 2) \\
\gamma &\sim \text{Uniform}(1, 4) \\
-(\alpha+1) &\sim \text{Beta}(-(\alpha+1) | 2, 2)
\end{aligned}
\tag{5.9}
$$

Equation 5.9 shows the details of our hierarchical Bayesian model, with the likelihood of the transit depth given the radius ratio and the transit depth measurement error in the first line; the interior structure models of Lop14 which map various planet

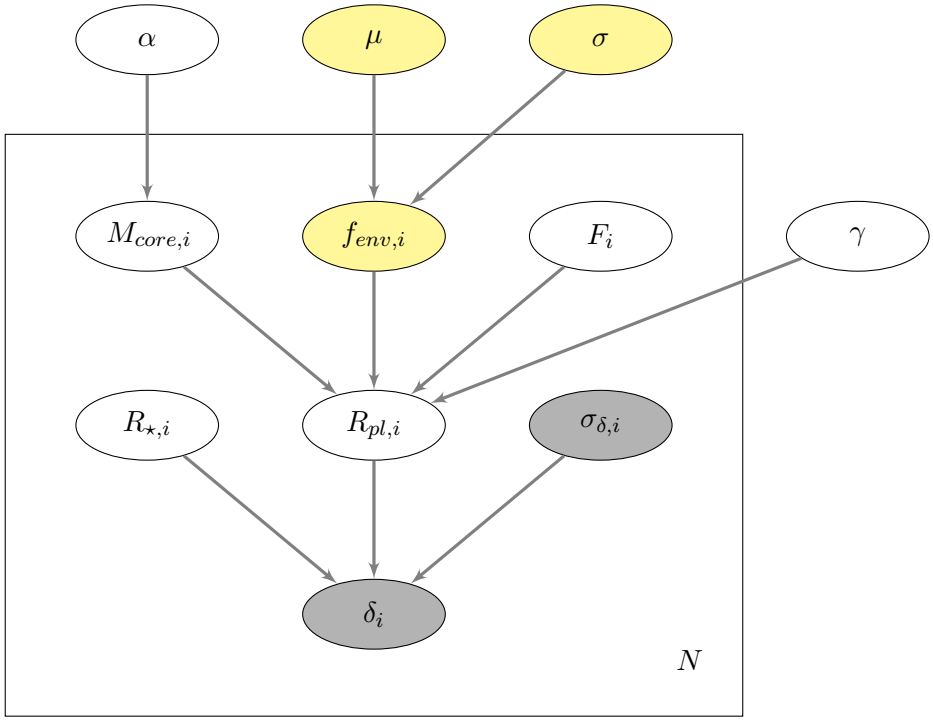

Figure 5.4: The graphical representation of our final hierarchical model (see Equation 5.9 for details). The hyperparameters, i.e. those which define population-wide distributions, are located outside the rectangle (called a "plate"), which represents the structure of individual parameters and data that is repeated for all of the planets in our sample ($i = 1, ..., N$=215). The yellow parameters are of interest in this work; they are constrained by the observed data (gray) through MCMC simulations (Section §5.3.5).

Data ($\boldsymbol{X}$):

δ_i = modeled transit depth

$\sigma_{\delta,i}$ = transit depth uncertainty

Parameters ($\boldsymbol{\theta}$):

$M_{core,i}$ = mass of Earth-like rocky core

$f_{env,i}$ = fraction of total mass in H+He envelope

F_i = incident stellar flux

$R_{pl,i}$ = planet radius

$R_{\star,i}$ = stellar radius

Hyperparameters ($\boldsymbol{\beta}$):

α = index of the $M_{core,i}$ power law distribution

μ = mean of the $f_{env,i}$ lognormal distribution

σ = standard deviation of the $f_{env,i}$ lognormal

γ = exponent of the envelope mass loss threshold

properties to radius in the second line; an analytic fit to the marginal Hub14 likelihood of each host star's radius in the third line; the priors on the individual planet property parameters in lines 4 - 5; and the priors on the hyperparameters in lines 6 - 9.

We have followed common practice and assumed a normal distribution for our likelihood, which is the equivalent assumption that one makes when using χ^2. Specifically this means that we have assumed that the measured transit depth δ_i is normally distributed around the "true" transit depth equal to the planet-star radius ratio squared, with standard deviation set by the error on the transit depth. The internal structure models are the power-law approximations given in Lop14; we did not use the full grid of models as the computational cost of interpolating a multi-dimensional grid was prohibitive within JAGS. Although this imposes a factor of ~ 2 theoretical uncertainty in our inferred f_{env} values, the width of the f_{env} posteriors is still dominated by the substantial radius uncertainties, and we proceed with the more computationally efficient choice. Due to similar concerns, we approximated the Hub14 stellar radius likelihoods that had been marginalized over all other stellar parameters as a gamma distribution by fitting its parameters a and b to each individual star.

As discussed above, we have assumed a lognormal distribution for the planets' gaseous envelope mass fractions (see §5.5.1 for an in-depth discussion of this choice), where μ and σ are the mean and standard deviation of the population of $\log(f_{env})$ values. For the planet core masses we also follow convention and use a power-law distribution, which is known as a Pareto distribution in statistics; it is parameterized by the power-law index α and a lower limit which we have set to 0.5 $M_\oplus$. We have truncated this

power law so that all $M_{core,i} < 20$ $M_\oplus$; this is motivated by both the work of Marcy et al. (2014), who find that planets in this size range have total masses between the mass detection threshold and 15-20 $M_\oplus$ (see Figure 49 of that paper), and the measurement of the most massive dense super-Earth found to date, Kepler-10c, at $M_{pl} \approx 17 \pm 2$ $M_\oplus$ (Dumusque et al., 2014).

As for the priors on the hyperparameters, we use a uniform distribution for the "location" parameter μ and a log-uniform distribution for the "scale" parameter σ^2. These distributions are equivalent to Jeffreys prior for these parameters and thus represent non-informative prior information (note that the uniform distribution is not always non-informative, especially for scale parameters or under parameter transformations). For the core mass power law index α, which must be > -1 for the power law to be proper[3], we use previous results from radial velocity surveys (i.e. Howard et al. 2010) and the intuition that smaller core masses must be more common to limit $0 < -(\alpha + 1) < 1$ with diffuse but higher probability density around 0.5 (an index transformation is needed due to the way statisticians define the Pareto distribution). This is naturally accomplished with the Beta distribution[4] whose parameters have both been set to 2. Finally, we allow for theoretical uncertainties in the evaporation threshold power law index by allowing γ to vary under a uniform prior distribution.

[3] A proper probability distribution cannot integrate to ∞ over its support.

[4] The Beta distribution is defined as $p_{Beta}(x|\alpha_B, \beta_B) \propto x^{\alpha_B - 1}(1-x)^{\beta_B - 1}$ so that $p_{Beta}(x|2,2). \propto x(1-x)$

5.3.5 JAGS: MCMC with Hierarchical Models

Having fully specified this hierarchical model and motivated our choices for specific distributions, we can now run Markov Chain Monte Carlo (MCMC) simulations to give us posteriors on all of our parameters of interest. Rather than write our own MCMC sampler, we use JAGS (Just Another Gibbs Sampler[5]; Plummer 2003), which was written specifically to analyze hierarchical Bayesian models via MCMC. Its platform independence and compatibility with the R computing language builds upon the BUGS project (Lunn et al., 2000), which was historically developed for analyzing hierarchical models on Windows platforms.

As its name suggests, JAGS uses Gibbs sampling to proceed from step to step in the Markov chain, which requires the ability to write down the full conditional probability distribution of each parameter. In practice, JAGS assigns different distributional families of Gibbs samplers to each parameter based on which sampling method is most efficient for the families of distributions involved in that region of the hierarchical model. Many full conditionals are algebraically complicated and become expensive to evaluate, in which case JAGS implements Adaptive Rejection Metropolis Sampling. The accepted parameter values are then stored and interpreted as samples from the target posterior distribution.

To produce the results shown in §5.4, we run our model with 10 chains, each for 500,000 iterations. The first half of each chain is discarded as "burn-in", and the resulting half is thinned by a factor of 250, such that we retain 10,000 posteriors samples

[5]JAGS code and user manuals can be downloaded at http://sourceforge.net/projects/mcmc-jags/.

of each parameter. JAGS computes the Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992) at run-time; the convergence of our MCMC simulations is analyzed in §5.4.3.

5.4 Results

Here we present the results of the hierarchical MCMC simulations for the parameters of interest in this work (highlighted yellow in Figure 5.4).

5.4.1 Population Composition Distribution

Figure 5.5 shows the results for the top-most level of our model (see §5.3.4): the population-wide composition parameters. More specifically, the left panel displays the marginal posterior distribution for μ and σ; these hyperparameters determine the mean and standard deviation, respectively, of the population distribution of $\log(f_{env})$ values. As f_{env} denotes the fraction of a planet's mass that exists in a hydrogen and helium envelope around an Earth-like rocky core, these parameters set the composition distribution of *Kepler*'s sub-Neptune planet candidates under our assumed internal structure. The "best-fit" μ and σ values, i.e. the mode of this posterior, are denoted by the large triangle and correspond to -2.2 dex ($\approx 0.7\%$) and 0.6 dex, respectively; they were found by performing two-dimensional kernel density estimation on a 50x50 grid and identifying the grid point with the highest density of posterior samples. This 2D KDE also gives us the drawn contours enclosing 68% and 95% of the posterior density. As the points in Figure 5.5 range over the allowed values of every parameter, utilizing

all of the posterior samples in the above calculation effectively marginalizes over all of the other parameters in Figure 5.4.

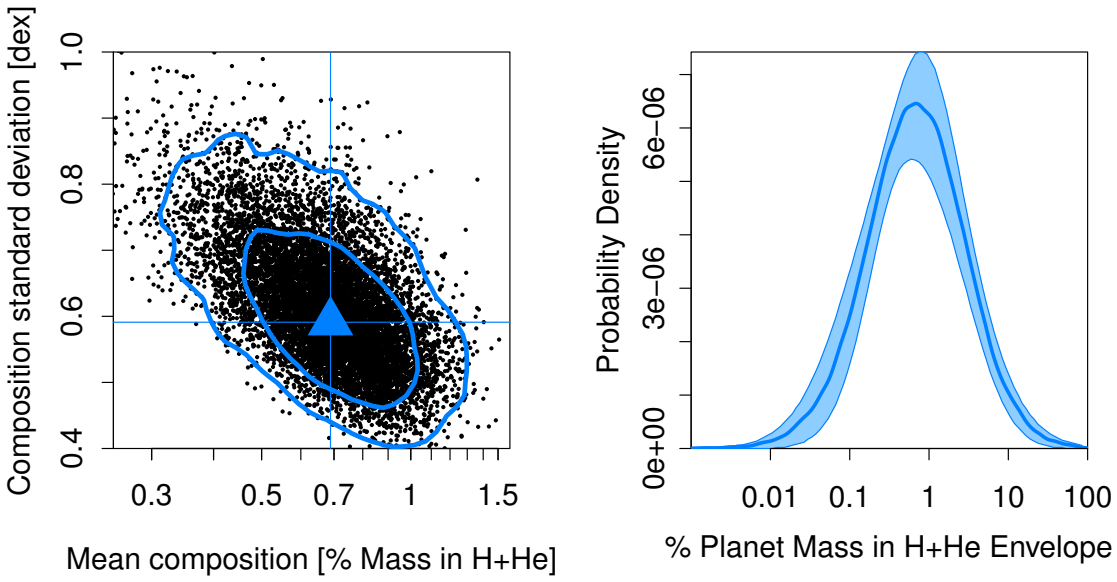

Figure 5.5: *Left:* Marginal posterior distribution for the mean μ and standard deviation σ of the $\log(f_{env})$ population distribution, with 68% and 95% contours. The "best-fit" mean log envelope fraction, denoted by the vertical line at the large triangle, is -2.2 ($f_{env} \approx 0.007$). *Right:* The posterior predictive composition distribution of *Kepler*'s sub-Neptune planet candidates (solid line), with a 68% coverage band. The peak of this f_{env} distribution corresponds to the "best-fit" value of μ, which shows that planets with $1\ R_\oplus < R_{pl} < 4\ R_\oplus$ and an incident stellar flux low enough to retain a gaseous envelope are most likely to be composed of $\sim 1\%$ H+He by mass.

To elucidate what the hyperparameter posterior implies for the sub-Neptune composition distribution, we must map the allowed (μ, σ) values onto f_{env} space. This is shown in the right panel of Figure 5.5, where we plot the posterior predictive distribution of $\log(f_{env})$ in solid blue. This distribution is computed by drawing (with replacement) 10,000 sets of (μ,σ) values from the posterior in the left panel of Figure 5.5, which defines 10,000 f_{env} distributions. From each of these we randomly draw one $\log(f_{env})$

value; combining all of these values into one histogram effectively marginalizes over the uncertainty in μ and σ and produces the $\log(f_{env})$ posterior predictive distribution. To compute the 68% coverage band in light blue[6], we draw several thousand $\log(f_{env})$ values from each set of (μ,σ), which results in 10,000 $\log(f_{env})$ histograms. On a bin-by-bin basis, we then find the range of counts which enclose 68% of the histograms. Note that this distribution does not include rocky planets; see Figures 5.6, 5.9, and 5.10 for discussion about the gas-rock transition. Additionally, $f_{env} \sim 0.1\%$ corresponds to a gaseous envelope that extends ~ 0.1 $R_\oplus$ above the rocky core, which is below the radius precision for these planets; the constraints on the distribution for the smallest f_{env} arise from the lognormal assumption.

The posterior predictive distribution (right panel of Figure 5.5) illustrates that the most likely composition for these sub-Neptune planet candidates is $\sim 1\%$ H+He by mass. This distribution represents the probability that, given the currently observed planet sample, the next observed planet in our considered size range (1-4 $R_\oplus$) will have a certain envelope fraction. It therefore marginalizes over planetary radius, meaning that the shape of the observed radius distribution for this complete subsample of planet candidates is encoded in the shape of the envelope fraction distribution. It is important to note that this distribution does not predict a planet's envelope fraction based on its measured radius; rather, it gives the distribution of envelope fractions over the entire population of sub-Neptunes. To see how well radius maps to composition for individual planets, see Figures 5.6 and 5.9.

[6]The coverage band is analogous to a confidence band in frequentist statistics, albeit with the requisite difference in interpretation, as it represents parameter uncertainty rather than variation between data sets.

5.4.2 Individual Planet Compositions

While Figure 5.5 gives the marginalized population distribution of compositions and so does not facilitate inferences that use knowledge of an individual's radius, our hierarchical model enabled us to compute individual composition posteriors for the 215 planet candidates in our complete *Kepler* subsample. These posteriors are summarized in Table 5.1. Matching an arbitrary *Kepler* planet's radius and radius uncertainty[7] to those given in Table 5.1 will give, to first order, the range of allowed compositions for that planet. Note that the radii given here are not exactly the same as the radii reported at the NExSci Exoplanet Archive, as the latter values do not use the full Hub14 stellar radius likelihood like we do here (also, see discussion about shrinkage in §5.4.3). Additionally, there will be some variation in composition for individual planets that are at different periods or that are hosted by stars of different spectral types, as these parameters do affect composition but to a lesser degree than radius. We include periods and stellar radii in Table 5.1 for these more detailed comparisons.

Figure 5.6 displays the information in Table 5.1, showing the individual planet compositions as a function of radius. Points denote the mode of the f_{env} and R_{pl} posteriors, while the lines denote the central 68% coverage interval (C.I.). If more than half of an individual's f_{env} posterior samples are zero, indicating that the stellar flux incident on the planet breached the mass loss threshold more than half of the time, we label that planet as "rocky" and give it a triangular symbol. For some of these planets, the 68% C.I. includes non-rocky compositions; these planets have f_{env} errors

[7]Uncertainties are expressed in terms of the coverage intervals which enclose the central 68% of the marginal posteriors (68% C.I.s).

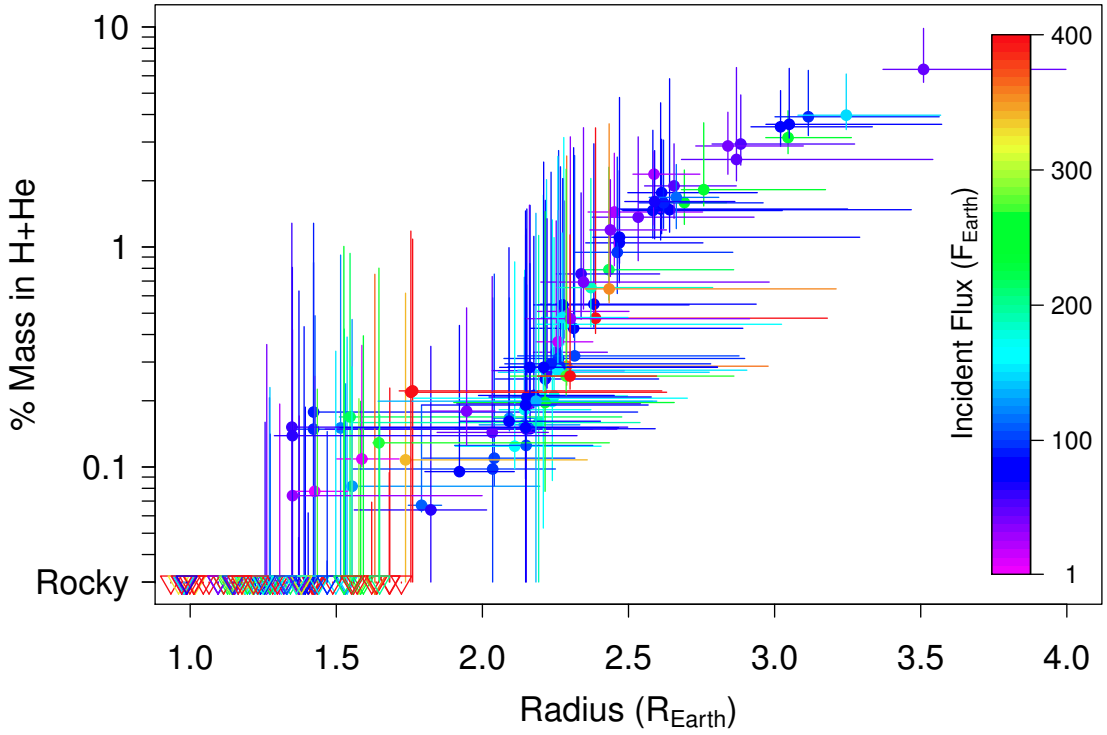

Figure 5.6: Individual planet compositions as a function of radius. Points denote the mode of the f_{env} and R_{pl} posteriors, while the lines denote the central 68% coverage intervals (C.I.s) of each parameter (the two-dimensional C.I.s are actually ellipses that are covariant along the direction of the R_{pl}, f_{env} locus, but for the sake of having all the points visible we choose to plot the marginal C.I.s as lines). Triangles denote rocky planets, for which more than half of the f_{env} posterior samples are zero. Color corresponds to the flux incident on the planet; we predict that the two planet candidates that have gaseous envelopes despite incident fluxes $F \geq 400$ $F_\oplus$ must have massive rocky cores, likely > 10 $M_\oplus$ (KOI 171.01 and KOI 355.01). There is a locus of allowed compositions and radii, such that planets with $R_{pl} < 2$ $R_\oplus$ have $f_{env} < 1\%$, planets with $2 < R_{pl} < 3$ $R_\oplus$ have $f_{env} \sim 1\%$, and planets with $R_{pl} > 3$ $R_\oplus$ have $f_{env} \sim$ a few %.

bars which extend into the gaseous region of parameter space. Color corresponds to the flux incident on the planet, given the period and the stellar parameters reported in Hub14; red points denote planets with $F \geq 400$ $F_\oplus$.

Table 5.1: Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
49.01	9527334	3.25	(3.08 , 3.57)	8.31	1.03	4.0	(3.4 , 6.1)
69.01	3544595	1.63	(1.59 , 1.69)	4.73	0.92	Rocky	(0 , 0)
70.01	6850504	3.05	(2.97 , 3.57)	10.85	0.92	3.6	(3.1 , 6.5)
70.02	6850504	1.74	(1.75 , 2.36)	3.70	0.93	0.1	(0.0 , 0.6)
70.05	6850504	0.98	(0.92 , 1.18)	19.58	0.91	Rocky	(0 , 0)
82.01	10187017	2.59	(2.52 , 2.74)	16.15	0.76	2.1	(1.2 , 2.7)
82.02	10187017	1.37	(1.34 , 1.43)	10.31	0.77	Rocky	(0 , 0)
84.01	2571238	2.47	(2.35 , 2.76)	9.29	0.86	1.0	(0.7 , 2.2)
103.01	2444412	2.88	(2.79 , 3.27)	14.91	0.91	2.9	(2.3 , 4.9)
104.01	10318874	2.66	(2.57 , 2.81)	2.51	0.64	1.7	(1.2 , 2.3)
112.02	10984090	1.30	(1.22 , 1.59)	3.71	1.06	Rocky	(0 , 0)
116.01	8395660	2.46	(2.32 , 2.86)	13.57	1.02	0.9	(0.6 , 2.5)
116.04	8395660	1.15	(1.06 , 1.34)	23.98	0.99	Rocky	(0 , 0)
139.02	8559644	1.57	(1.45 , 1.82)	3.34	1.12	Rocky	(0 , 0)
148.01	5735762	2.15	(2.06 , 2.37)	4.78	0.90	0.2	(0.1 , 0.8)
148.02	5735762	3.02	(2.92 , 3.33)	9.67	0.88	3.5	(2.8 , 5.1)
150.01	7626506	2.28	(2.19 , 3.02)	8.41	0.85	0.4	(0.3 , 3.1)
153.01	12252424	2.44	(2.37 , 2.63)	8.93	0.71	1.2	(0.7 , 2.0)
153.02	12252424	2.16	(2.09 , 2.32)	4.75	0.71	0.3	(0.2 , 0.8)
157.01	6541920	3.12	(3.00 , 3.56)	13.02	1.09	3.9	(3.2 , 6.3)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
157.02	6541920	3.51	(3.37 , 4.00)	22.69	1.09	6.4	(5.6 , 9.8)
157.06	6541920	2.14	(1.99 , 2.33)	10.30	1.07	0.2	(0.1 , 0.7)
159.01	8972058	2.37	(2.26 , 2.79)	8.99	1.05	0.7	(0.4 , 2.0)
159.02	8972058	0.99	(0.93 , 1.24)	2.40	1.04	Rocky	(0 , 0)
161.01	5084942	2.69	(2.63 , 2.84)	3.11	0.82	1.6	(1.2 , 2.2)
162.01	8107380	2.64	(2.48 , 3.47)	14.01	1.08	1.5	(1.1 , 5.8)
165.01	9527915	2.66	(2.56 , 2.87)	13.22	0.81	1.9	(1.3 , 2.9)
166.01	2441495	2.25	(2.18 , 2.43)	12.49	0.77	0.3	(0.3 , 1.3)
167.01	11666881	1.76	(1.74 , 2.63)	4.92	1.19	0.2	(0.0 , 1.2)
171.01	7831264	2.39	(2.29 , 3.18)	5.97	1.18	0.5	(0.4 , 3.4)
171.02	7831264	2.21	(1.76 , 2.70)	13.07	1.16	0.2	(0.0 , 1.7)
172.01	8692861	2.28	(2.18 , 2.71)	13.72	0.90	0.5	(0.3 , 2.0)
177.01	6803202	2.22	(2.03 , 2.49)	21.06	1.14	0.3	(0.2 , 1.3)
180.01	9573539	2.62	(2.54 , 2.96)	10.05	0.94	1.6	(1.2 , 3.0)
238.01	7219825	2.58	(2.48 , 3.03)	17.23	1.11	1.5	(1.1 , 3.4)
273.01	3102384	2.09	(2.04 , 2.21)	10.57	1.08	0.2	(0.1 , 0.6)
280.01	4141376	2.15	(2.09 , 2.26)	11.87	1.04	0.2	(0.1 , 0.6)
282.02	5088536	1.11	(1.08 , 1.19)	8.46	1.13	Rocky	(0 , 0)
283.01	5695396	2.34	(2.25 , 2.61)	16.09	1.03	0.8	(0.4 , 1.7)
299.01	2692377	1.42	(1.36 , 1.59)	1.54	0.94	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
305.01	6063220	1.79	(1.75 , 1.86)	4.60	0.76	0.07	(0.0 , 0.2)
306.01	6071903	2.45	(2.36 , 2.75)	24.31	0.87	1.4	(0.8 , 2.6)
307.01	6289257	1.82	(1.56 , 2.01)	19.67	1.06	0.06	(0.0 , 0.3)
307.02	6289257	1.19	(1.11 , 1.33)	5.21	1.04	Rocky	(0 , 0)
312.01	7050989	2.11	(1.91 , 2.41)	11.58	1.16	0.1	(0.1 , 0.8)
312.02	7050989	2.15	(1.91 , 2.38)	16.40	1.18	0.1	(0.1 , 0.8)
313.01	7419318	2.28	(2.19 , 2.50)	18.74	0.86	0.5	(0.4 , 1.6)
313.02	7419318	1.92	(1.80 , 2.11)	8.44	0.86	0.1	(0.1 , 0.4)
314.01	7603200	1.59	(1.50 , 1.72)	13.78	0.51	0.1	(0.1 , 0.3)
314.02	7603200	1.43	(1.35 , 1.54)	23.09	0.52	0.08	(0.0 , 0.2)
321.01	8753657	1.42	(1.33 , 1.62)	2.43	1.03	Rocky	(0 , 0)
323.01	9139084	2.27	(2.20 , 2.50)	5.84	0.89	0.5	(0.3 , 1.1)
327.01	9881662	1.56	(1.48 , 1.71)	3.25	1.11	Rocky	(0 , 0)
333.01	10337258	2.26	(2.12 , 2.91)	13.29	1.14	0.3	(0.2 , 2.6)
352.02	11521793	2.15	(1.79 , 2.57)	16.01	0.99	0.2	(0.1 , 1.4)
354.01	11568987	2.59	(2.49 , 2.86)	15.96	1.04	1.6	(1.1 , 2.7)
354.02	11568987	1.25	(1.19 , 1.39)	7.38	1.03	Rocky	(0 , 0)
355.01	11621223	2.30	(2.19 , 2.60)	4.90	1.13	0.3	(0.2 , 1.1)
361.01	12404954	1.55	(1.48 , 1.76)	3.25	0.98	Rocky	(0 , 0)
369.01	7175184	1.32	(1.18 , 1.71)	5.89	1.16	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
369.02	7175184	1.33	(1.15 , 1.67)	10.10	1.14	Rocky	(0 , 0)
385.01	3446746	2.16	(1.91 , 2.59)	13.15	1.00	0.1	(0.1 , 1.5)
409.01	5444548	2.61	(2.46 , 3.25)	13.25	1.04	1.5	(1.0 , 4.5)
568.01	7595157	1.58	(1.44 , 2.11)	3.38	0.89	Rocky	(0.0 , 0.2)
568.02	7595157	1.05	(0.96 , 1.38)	2.36	0.87	Rocky	(0 , 0)
623.01	12068975	1.36	(1.31 , 1.41)	10.35	1.11	Rocky	(0 , 0)
623.02	12068975	1.33	(1.29 , 1.39)	15.68	1.11	Rocky	(0 , 0)
623.03	12068975	1.16	(1.14 , 1.23)	5.60	1.11	Rocky	(0 , 0)
627.01	4563268	2.43	(2.34 , 2.86)	7.75	1.17	0.8	(0.5 , 2.3)
627.02	4563268	1.42	(1.31 , 1.64)	4.17	1.16	Rocky	(0 , 0)
628.01	4644604	2.22	(2.05 , 2.60)	14.49	0.97	0.3	(0.2 , 1.6)
632.01	4827723	1.53	(1.49 , 2.01)	7.24	0.89	Rocky	(0.0 , 0.2)
639.01	5120087	2.38	(2.20 , 2.94)	17.98	1.14	0.5	(0.4 , 2.9)
647.01	5531694	1.68	(1.46 , 2.19)	5.17	1.11	Rocky	(0.0 , 0.2)
650.01	5786676	2.53	(2.38 , 2.93)	11.96	0.81	1.4	(0.8 , 3.1)
662.01	6365156	2.25	(2.10 , 2.50)	10.21	1.16	0.2	(0.2 , 1.1)
664.01	6442340	2.04	(1.55 , 2.25)	13.14	1.05	0.1	(0.0 , 0.6)
664.02	6442340	1.20	(1.11 , 1.45)	7.78	1.04	Rocky	(0 , 0)
664.03	6442340	1.09	(1.01 , 1.30)	23.44	1.04	Rocky	(0 , 0)
665.01	6685609	2.29	(2.15 , 2.98)	5.87	1.10	0.3	(0.2 , 2.6)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
665.02	6685609	1.15	(1.05 , 1.75)	1.61	1.11	Rocky	(0 , 0)
665.03	6685609	1.15	(1.02 , 1.68)	3.07	1.11	Rocky	(0 , 0)
666.01	6707835	2.84	(2.73 , 3.10)	22.25	1.04	2.9	(2.1 , 4.1)
673.01	7124613	1.76	(1.72 , 2.61)	4.42	1.14	0.2	(0.0 , 1.1)
691.02	8480285	1.24	(1.12 , 1.44)	16.23	1.02	Rocky	(0 , 0)
692.01	8557374	1.43	(1.29 , 1.78)	2.46	0.98	Rocky	(0 , 0)
692.02	8557374	1.65	(1.63 , 2.43)	4.82	0.97	0.1	(0.0 , 0.8)
693.02	8738735	2.27	(2.09 , 2.80)	15.66	1.10	0.3	(0.2 , 2.3)
694.01	8802165	2.87	(2.68 , 3.54)	17.42	0.94	2.5	(2.0 , 6.5)
700.02	8962094	1.45	(1.41 , 2.14)	9.36	0.91	Rocky	(0.0 , 0.3)
700.03	8962094	1.39	(1.30 , 1.94)	14.67	0.93	Rocky	(0.0 , 0.2)
701.01	9002278	2.26	(2.18 , 2.38)	18.16	0.66	0.4	(0.3 , 1.4)
701.02	9002278	1.47	(1.44 , 1.56)	5.71	0.66	Rocky	(0 , 0)
704.01	9266431	2.35	(2.20 , 2.98)	18.40	0.91	0.7	(0.5 , 3.5)
708.01	9530945	2.47	(2.38 , 3.29)	17.41	1.08	1.1	(0.8 , 4.8)
708.02	9530945	2.21	(1.90 , 2.66)	7.69	1.11	0.2	(0.1 , 1.4)
709.01	9578686	2.30	(2.15 , 2.92)	21.39	0.89	0.5	(0.4 , 3.1)
711.02	9597345	1.64	(1.49 , 1.82)	3.62	1.04	Rocky	(0 , 0)
714.01	9702072	2.76	(2.68 , 3.17)	4.18	0.88	1.8	(1.5 , 3.6)
717.01	9873254	2.09	(1.92 , 2.40)	14.71	1.11	0.2	(0.1 , 1.0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
719.01	9950612	1.95	(1.83 , 2.04)	9.03	0.71	0.2	(0.1 , 0.5)
984.01	1161345	3.05	(2.97 , 3.26)	4.29	0.91	3.1	(2.6 , 4.1)
987.01	7295235	1.40	(1.34 , 1.60)	3.18	0.92	Rocky	(0 , 0)
1002.01	1865042	1.30	(1.21 , 1.58)	3.48	1.00	Rocky	(0 , 0)
1116.01	2849805	1.68	(1.49 , 2.03)	3.75	1.14	Rocky	(0.0 , 0.1)
1118.01	2853446	1.58	(1.49 , 2.05)	7.37	1.02	Rocky	(0.0 , 0.2)
1128.01	6362874	1.15	(1.10 , 1.30)	0.98	0.88	Rocky	(0 , 0)
1150.01	8278371	1.12	(1.05 , 1.51)	0.68	1.16	Rocky	(0 , 0)
1151.01	8280511	1.39	(1.29 , 1.53)	10.44	0.87	Rocky	(0 , 0)
1165.01	10337517	2.29	(2.10 , 2.86)	7.05	0.94	0.3	(0.2 , 2.2)
1216.01	3839488	1.54	(1.49 , 2.17)	11.13	1.07	Rocky	(0.0 , 0.4)
1245.01	6693640	2.22	(2.05 , 2.78)	13.72	1.16	0.3	(0.2 , 2.0)
1279.01	8628758	2.18	(1.99 , 2.45)	14.37	1.00	0.2	(0.2 , 1.1)
1279.02	8628758	1.16	(1.08 , 1.41)	9.65	1.03	Rocky	(0 , 0)
1315.01	10928043	1.55	(1.45 , 1.69)	6.85	1.14	Rocky	(0 , 0)
1379.01	7211221	1.29	(1.22 , 1.54)	5.62	0.88	Rocky	(0 , 0)
1438.01	11193263	1.53	(1.36 , 2.24)	6.91	1.05	Rocky	(0.0 , 0.4)
1529.01	9821454	2.15	(1.51 , 2.47)	17.98	1.10	0.2	(0.0 , 1.1)
1529.02	9821454	1.23	(1.10 , 1.62)	11.87	1.08	Rocky	(0 , 0)
1531.01	11764462	1.38	(1.24 , 1.87)	5.70	1.02	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
1533.01	7808587	1.55	(1.32 , 2.06)	6.24	1.09	Rocky	(0.0 , 0.1)
1534.01	4741126	1.42	(1.42 , 2.53)	20.42	1.20	0.2	(0.0 , 1.3)
1534.02	4741126	1.02	(0.97 , 1.62)	7.64	1.13	Rocky	(0 , 0)
1606.01	9886661	1.65	(1.62 , 1.96)	5.08	0.94	Rocky	(0.0 , 0.1)
1608.01	10055126	1.55	(1.48 , 2.11)	9.18	1.05	Rocky	(0.0 , 0.3)
1608.02	10055126	1.37	(1.26 , 1.59)	19.74	1.06	Rocky	(0 , 0)
1628.01	6975129	2.61	(2.50 , 2.94)	19.75	1.13	1.8	(1.1 , 3.1)
1629.01	8685497	1.43	(1.31 , 1.71)	4.41	1.15	Rocky	(0 , 0)
1632.01	9277896	1.37	(1.13 , 1.73)	4.59	1.15	Rocky	(0 , 0)
1738.01	4365645	1.13	(1.07 , 1.48)	4.17	0.80	Rocky	(0 , 0)
1792.03	8552719	1.33	(1.26 , 1.50)	9.11	1.03	Rocky	(0 , 0)
1802.01	11298298	2.43	(2.35 , 3.21)	5.25	1.09	0.6	(0.5 , 3.6)
1806.02	9529744	1.39	(1.25 , 2.18)	17.93	1.17	Rocky	(0.0 , 0.4)
1806.03	9529744	1.20	(1.02 , 1.58)	8.37	1.12	Rocky	(0 , 0)
1809.01	8240797	2.32	(2.12 , 2.88)	13.09	1.17	0.3	(0.3 , 2.6)
1809.02	8240797	1.63	(1.53 , 2.43)	4.92	1.18	Rocky	(0.0 , 0.7)
1819.01	9597058	2.03	(1.85 , 2.23)	12.06	0.73	0.1	(0.1 , 0.7)
1820.01	8277797	1.53	(1.45 , 2.52)	4.34	0.82	Rocky	(0 , 1)
1837.02	10657406	1.22	(1.09 , 1.70)	1.68	0.94	Rocky	(0 , 0)
1850.01	8826168	2.15	(2.02 , 2.58)	11.55	0.97	0.2	(0.2 , 1.5)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
1886.01	9549648	1.64	(1.51 , 1.78)	5.99	1.12	Rocky	(0 , 0)
1893.01	8689793	1.62	(1.42 , 1.96)	3.56	0.97	Rocky	(0 , 0)
1898.01	7668663	1.59	(1.49 , 2.24)	6.50	1.14	Rocky	(0.0 , 0.4)
1899.01	7047922	2.26	(2.07 , 2.90)	19.76	1.14	0.3	(0.2 , 2.7)
1909.01	10130039	1.47	(1.40 , 1.85)	12.76	1.02	Rocky	(0.0 , 0.1)
1909.02	10130039	1.14	(1.08 , 1.38)	5.47	1.01	Rocky	(0 , 0)
1913.01	9704384	1.44	(1.38 , 1.61)	5.51	0.95	Rocky	(0 , 0)
1916.01	6037581	2.16	(1.91 , 2.54)	20.68	0.99	0.2	(0.2 , 1.5)
1916.02	6037581	1.51	(1.51 , 2.42)	9.60	0.99	0.2	(0.0 , 0.9)
1937.01	10190777	1.21	(1.15 , 1.29)	1.41	0.61	Rocky	(0 , 0)
1955.01	9892816	2.18	(1.64 , 2.60)	15.17	1.17	0.2	(0.0 , 1.4)
1960.01	6949061	2.19	(1.56 , 2.54)	8.97	1.13	0.2	(0.0 , 1.1)
1960.02	6949061	2.15	(1.46 , 2.46)	23.22	1.05	0.1	(0.0 , 1.2)
1963.01	10917681	2.23	(2.08 , 2.78)	12.90	1.02	0.3	(0.2 , 2.2)
1972.01	11253711	2.21	(2.06 , 2.80)	17.79	1.06	0.3	(0.2 , 2.4)
1979.01	7273277	1.00	(0.96 , 1.52)	2.71	0.75	Rocky	(0 , 0)
2007.02	11069176	1.35	(1.29 , 2.32)	21.13	1.07	0.1	(0.0 , 0.8)
2011.01	5384079	1.37	(1.19 , 1.86)	7.06	1.15	Rocky	(0 , 0)
2011.02	5384079	1.10	(0.99 , 1.57)	17.27	1.21	Rocky	(0 , 0)
2017.01	8750043	1.28	(1.14 , 1.72)	2.30	0.87	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
2026.01	11923284	1.72	(1.50 , 2.03)	2.76	1.12	Rocky	(0 , 0)
2029.01	9489524	1.35	(1.37 , 2.00)	16.33	0.82	0.07	(0.0 , 0.4)
2032.01	2985767	1.27	(1.14 , 1.94)	14.08	0.91	Rocky	(0.0 , 0.2)
2033.01	2304320	1.31	(1.28 , 1.71)	16.54	0.67	Rocky	(0.0 , 0.2)
2049.01	9649706	1.49	(1.31 , 1.93)	1.57	1.12	Rocky	(0 , 0)
2053.01	2307415	1.55	(1.56 , 2.20)	13.12	1.09	0.08	(0.0 , 0.4)
2053.02	2307415	1.41	(1.30 , 1.64)	4.61	1.11	Rocky	(0 , 0)
2059.01	12301181	0.98	(0.95 , 1.06)	6.15	0.79	Rocky	(0 , 0)
2087.01	6922710	1.37	(1.30 , 1.83)	23.13	1.05	Rocky	(0.0 , 0.1)
2105.01	8165946	1.44	(1.32 , 2.15)	6.42	1.07	Rocky	(0.0 , 0.2)
2110.01	11460462	1.01	(0.99 , 1.73)	5.04	1.16	Rocky	(0 , 0)
2137.01	9364609	1.35	(1.36 , 2.50)	14.97	0.91	0.2	(0.0 , 1.3)
2159.01	8804455	1.26	(1.15 , 1.50)	7.60	1.01	Rocky	(0 , 0)
2246.01	9458343	1.43	(1.31 , 2.23)	11.90	1.05	Rocky	(0.0 , 0.5)
2278.01	3342794	2.04	(1.80 , 2.32)	14.17	1.03	0.1	(0.1 , 0.7)
2278.02	3342794	1.01	(0.94 , 1.26)	4.92	1.03	Rocky	(0 , 0)
2281.01	9221517	0.97	(0.90 , 1.23)	0.77	0.84	Rocky	(0 , 0)
2331.01	12401863	1.23	(1.12 , 1.74)	2.83	1.09	Rocky	(0 , 0)
2333.01	11121752	1.18	(1.10 , 1.44)	3.93	1.07	Rocky	(0 , 0)
2333.02	11121752	1.38	(1.13 , 1.54)	7.63	1.06	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
2342.01	10212441	1.15	(1.04 , 1.37)	15.04	1.00	Rocky	(0 , 0)
2389.01	8494617	1.32	(1.23 , 1.49)	22.92	1.02	Rocky	(0 , 0)
2403.01	2142522	1.20	(1.10 , 1.60)	13.32	1.07	Rocky	(0 , 0)
2414.01	8611832	1.10	(1.03 , 1.30)	22.60	0.85	Rocky	(0 , 0)
2443.01	9209624	1.26	(1.07 , 1.61)	6.79	1.11	Rocky	(0 , 0)
2443.02	9209624	1.29	(1.06 , 1.56)	11.84	1.11	Rocky	(0 , 0)
2555.01	5350244	1.24	(1.05 , 1.54)	12.57	1.17	Rocky	(0 , 0)
2559.01	6605493	1.31	(1.21 , 1.52)	9.31	1.10	Rocky	(0 , 0)
2563.01	5175024	1.37	(1.28 , 2.27)	23.48	1.16	Rocky	(0.0 , 0.6)
2675.01	5794570	2.24	(1.91 , 2.58)	5.45	0.85	0.2	(0.1 , 1.3)
2693.03	5185897	0.98	(0.92 , 1.07)	6.83	0.68	Rocky	(0 , 0)
2711.01	5272233	1.55	(1.51 , 2.48)	9.02	1.13	0.2	(0.0 , 0.9)
2711.02	5272233	1.42	(1.36 , 2.38)	17.34	1.09	0.1	(0.0 , 0.8)
2730.01	8415200	1.18	(1.05 , 1.61)	4.52	1.06	Rocky	(0 , 0)
2732.01	9886361	1.13	(1.08 , 1.35)	7.03	1.15	Rocky	(0 , 0)
2732.02	9886361	1.25	(1.16 , 1.43)	13.61	1.15	Rocky	(0 , 0)
2743.01	8095441	1.26	(1.17 , 1.88)	11.88	0.82	Rocky	(0.0 , 0.1)
2906.01	6716545	1.19	(1.04 , 1.51)	13.91	1.13	Rocky	(0 , 0)
2971.01	4770174	0.96	(0.93 , 1.48)	6.10	1.14	Rocky	(0 , 0)
2984.01	7918652	0.98	(0.95 , 1.42)	11.46	1.13	Rocky	(0 , 0)

Table 5.1 (cont'd): Compositions of Individual Sub-Neptune-sized Planets in Sample

KOI #	Kepler ID	R_{pl} ($R_\oplus$)	68% C.I. ($R_\oplus$)	P (days)	$R_\star$ ($R_\odot$)	f_{env} (%)	68% C.I. (%)
3020.01	8241079	1.22	(1.03 , 1.57)	10.92	1.13	Rocky	(0 , 0)
3075.01	3328080	0.93	(0.93 , 1.49)	4.77	0.98	Rocky	(0 , 0)
3209.01	7017274	1.34	(1.18 , 1.71)	11.91	1.12	Rocky	(0 , 0)
3301.01	8301878	1.26	(1.14 , 2.05)	20.71	0.97	Rocky	(0.0 , 0.3)
3346.01	11241912	1.26	(1.15 , 1.44)	14.43	1.05	Rocky	(0 , 0)
3384.01	8644365	1.22	(1.12 , 1.43)	10.55	1.12	Rocky	(0 , 0)
3384.02	8644365	1.40	(1.31 , 1.64)	19.92	1.11	Rocky	(0 , 0)
3438.01	6599975	1.27	(1.12 , 2.08)	14.56	1.17	Rocky	(0.0 , 0.2)
3876.01	3440118	2.31	(2.16 , 2.89)	19.58	1.16	0.4	(0.4 , 2.8)
3880.01	4147444	1.15	(1.00 , 1.68)	1.80	1.12	Rocky	(0 , 0)
4022.01	7733731	1.06	(0.95 , 1.44)	4.86	1.06	Rocky	(0 , 0)
4053.01	1718958	0.98	(0.95 , 1.58)	1.42	1.10	Rocky	(0 , 0)
4320.01	5095082	0.97	(0.91 , 1.28)	20.66	0.86	Rocky	(0 , 0)
4335.01	10730070	1.27	(1.10 , 1.70)	7.62	1.08	Rocky	(0 , 0)
4505.01	8493354	1.27	(1.04 , 1.94)	18.01	1.20	Rocky	(0.0 , 0.1)

Note. — The reported R_{pl} is the peak of the marginal posterior planet radius distribution, and the 68% C.I. is the coverage interval which encloses its central 68% probability region, which is dominated by the stellar radius uncertainties. Note that these are not exactly the same as the radii reported at the NExSci Exoplanet Archive, as those values do not use the full Hub14 stellar radius likelihood like we do here. Furthermore, the two-dimensional C.I.s are actually ellipses that are covariant along the direction of the R_{pl}, f_{env} locus shown in Figure 5.6; for reporting simplicity, the marginal C.I.s are given here.

5.4.3 Posterior Checks and Convergence

An important part of Bayesian analysis is testing for the convergence of the MCMC simulations, which we check for in a number of ways. To begin, we compare the prior distributions for individual planet parameters to their posteriors for a quick yet illustrative reality check that our hierarchical MCMC simulations are producing reasonable results. If we have strong prior information about the parameters, then the hyperparameter posteriors and the structure of the statistical model should preserve this information via posteriors that are similar in shape and location to the priors.

Figure 5.7 shows this check for the planet radii, which we treat as a parameter

in our model and have strong prior information for (see Figure 5.4 and Equation 5.9). On the x-axis we plot the "prior" radius distribution for each planet in our sample[8], which we compute by scaling the Hub 14 stellar radius likelihoods by the observed transit depth, and on the y-axis we plot the posterior radius distribution that result from our MCMC simulations. The modes of the distributions are denoted as points, with the 68% coverage interval spanned by the lines. The color of the points denote the value of the Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992) for each planet's f_{env} posterior, which we discuss in greater detail below. The diagonal green line is the 45° line, which we expect all of our individual radius distributions to span if our model is behaving as required. We immediately see that this is the case, indicating that our model is incorporating our prior radius information appropriately and that our posteriors are accurate given our data and its assumed hierarchical structure.

We also see a few salient features of our model manifest in this figure. First, there is a zone of avoidance between 1.7 and 2.0 $R_\oplus$ where the posterior radius distributions have been pushed to either side of the corresponding prior distributions (but not unreasonably so, given that each planet has a 68% coverage interval spans the one-to-one line). This is due to our incorporation of photoevaporation, as with periods < 25 days it is difficult to retain the tenuous gaseous envelope needed to produce these planetary radii. Note that the planet radius priors are wide enough that this model feature does not pose significant problems for inferring a H+He envelope composition for each planet; however, if the planet parameters were more tightly constrained, as is the case for those

[8]The input planet radius ($R_{pl,i}$) distributions plotted here are not priors in the strictest sense of the definition, as the stellar radius is actually the quantity that has a distribution determined *a priori*; however, since the transit depth uncertainties are small, the prior $R_{pl,i}$ distributions can be reasonably approximated as described here.

orbiting brighter host stars with spectroscopic follow-up, a radius falling solidly within this narrow range would be better explained with a water-dominated composition, as we discuss in §5.1.1.

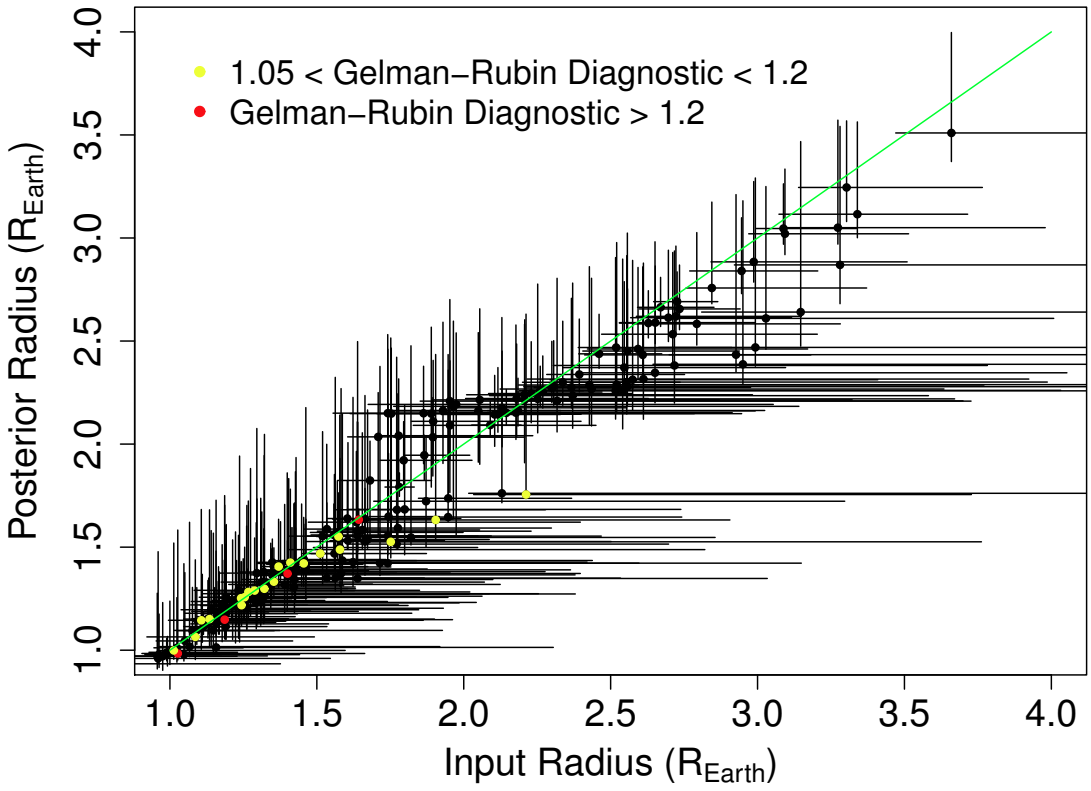

Figure 5.7: Comparison of the "prior" planet radius distributions to the posterior distributions, with convergence diagnostics for the corresponding f_{env} posterior. The 45° line where posterior estimates equal prior estimates is green. Each planet has a 68% coverage interval that spans the one-to-one line, indicating that our model is behaving appropriately.

Second, a general feature of hierarchical Bayesian models is evident in Figure 5.7: the posteriors on these individual parameters are much narrower than the priors. While this is expected for all Bayesian analyses given the role of data in constraining

posterior estimates from prior information, the hierarchical structure of our problem contributes to this effect: by relating individuals to each other, HBM provides posterior estimates of individual exoplanet properties which have smaller variance than if multiple individual Bayesian analyses were performed independently. This is called shrinkage, as this effect is achieved by "shrinking" the posterior estimates toward the population mean (see Loredo 2007 for a more detailed discussion). Indeed, Figure 5.7 shows that the points above 2.5 $R_\oplus$ fall slightly below the one-to-one line, and the points below 2 $R_\oplus$ fall slightly above it, as the mean envelope fraction of about 1% roughly corresponds to a radius of ~ 2.2 $R_\oplus$.

We continue the discussion of convergence with Gelman-Rubin convergence statistic ($\hat{R}$) for the individual f_{env} posteriors. This diagnostic calculates the ratio of the total variance across all chains in the MCMC simulation to the variance within individual chains; $\hat{R}$ within a percent or so of 1 indicates convergence, where each individual chain probes about the same volume of parameter space as all of the chains taken together. Most of the individual f_{env} posteriors have $\hat{R} \leq 1.01$, but there are some planets whose $\hat{R}$ values indicate that the MCMC should be run longer. One immediately notices that these planets have small radii; in fact, every planet with $\hat{R} > 1.01$ has a f_{env} posterior that spans 0. Given the discrete nature of the switch between rocky and gaseous compositions, the fact that mixing between f_{env} chains is worse for the planets which cross this transition is not a surprise; additionally, we expect $\hat{R}$ to be biased high simply as a numerical artifact of representing rocky compositions with $f_{env} = 0$, as this imposes a gap between rocky and non-rocky f_{env} chains. Noting that the planets with

rocky compositions do not contribute to constraints on the composition distribution hyperparameters, we conclude that these $\hat{R}$ values are not a cause for concern.

Similarly, we must assess the convergence of the composition hyperparameters in our simulation. The $\hat{R}$ values for μ and σ are 1.08 and 1.03, respectively. However, as we see above, $\hat{R}$ does not always convey the full picture of convergence, so we turn to other common diagnostics such as trace plots and autocorrelation functions (Figure 5.8, top and bottom rows respectively) to more fully investigate the issue. In the trace plots, the values of a parameter's chain is plotted as a function of location along the chain, with different colors indicating different chains. We see that there is good mixing between the chains for both parameters, indicating that we have arrived at the stationary distribution for the joint posterior displayed in Figure 5.5.

The average chain autocorrelation functions further support the convergence of our simulations, as they quickly reach a low level of autocorrelation. The slightly higher autocorrelation present in μ explains the slightly higher $\hat{R}$ calculated for that parameter, but the difference is not strong enough to be visible in the trace plots. Given that the mode of the μ posterior distribution has been well established through the mixing of the existing chains, our conclusion that the most likely sub-Neptune composition is $\sim 1\%$ H+He by mass would not change by running the simulation longer. With such diminishing returns regarding convergence, we take our (μ,σ) posterior as the stationary distribution and continue with a discussion of these results.

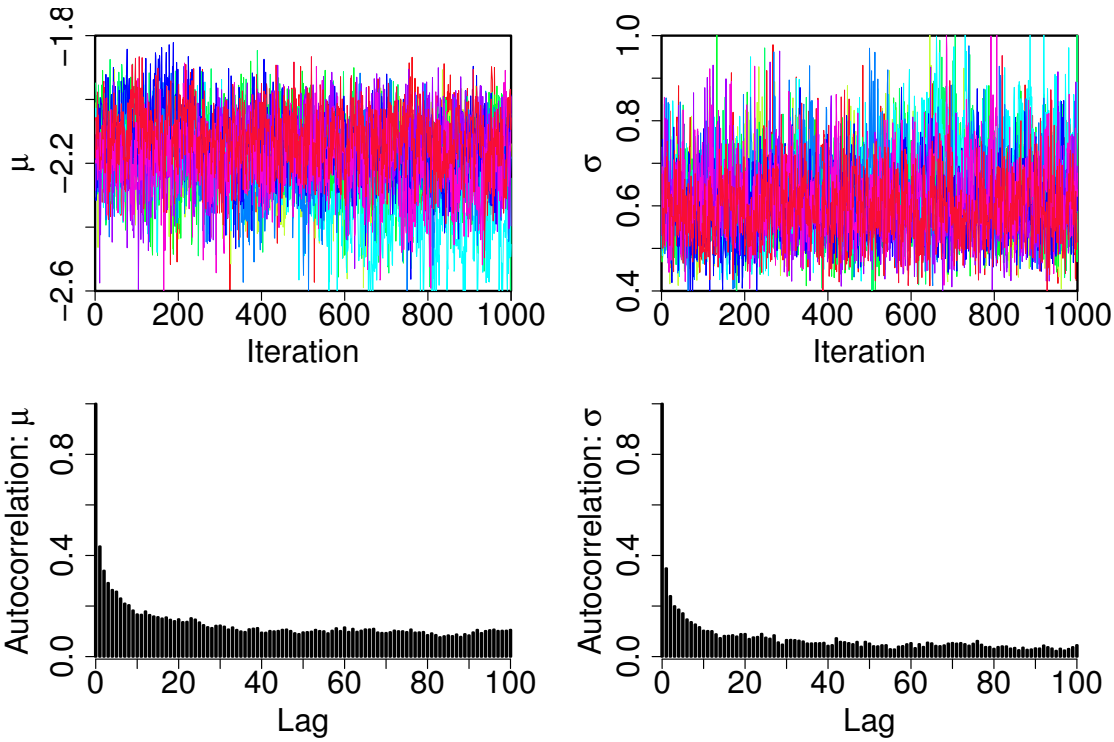

Figure 5.8: Graphical convergence diagnostics for our composition hyperparameters μ (left column) and σ (right column); the first row displays the value of all 10 MCMC chains as a function of location in the chain (a "trace plot"), while the bottom row displays the average chain autocorrelation at increasing offsets. Both parameters have good mixing between the chains and quickly reach a low level of autocorrelation, indicating that we have converged to the stationary distribution for the joint posterior displayed in Figure 5.5.

5.5 Discussion

The results in §5.4 have numerous implications for characterizing the sizable sub-Neptune population discovered by *Kepler*. We discuss several in detail below, but first we address some of the more constraining choices that we have made in our statistical model (see §5.3.2 - 5.3.4 for further description and motivation for all of the assumptions we make).

5.5.1 Motivation for Salient Model Assumptions

Arguably the most obvious assumption we've made other than the rock/H+He interior structure is that the composition distribution can be reasonably described with a lognormal distribution. There is currently very little theoretical guidance regarding the expected shape of this distribution; in the absence of such predictions, we turn to our driving science questions (see §5.1) to inform this choice. Because we are interested in characterizing the "typical" sub-Neptune envelope fraction as well as the range present in the population, the most natural choice is a distribution that straightforwardly parameterizes the mean and variance of a population: a normal distribution. In addition, we expected a large dynamic range of gaseous envelope fractions, which the lognormal in particular is able to accommodate. We acknowledge that different choices for this composition distribution can affect the result we present here, as the particular parametric form drives the quantitative details of the shrinkage we observe in §5.4.3. Alternatively, one could completely sidestep this concern by adopting a nonparametric approach; however, doing so involves solving for a much larger number of free parameters, which simultaneously reduces the predictive power and expands the computational expense of such a study. Our choice therefore best balances the current demands of our scientific goals, our computational considerations, and the desire to limit the number of free parameters in an already fairly complex statistical model. That said, there is no reason why one could not assume a mixture of lognormals or any other more flexible distribution in the future with a larger planet sample and more computing power.

We also make several assumptions that are not explicit in our statistical model.

First, we assume that all of the planet candidates in our sample are true planets. If we were concerned with an absolute occurrence rate of planet compositions, we would need to correctly account for the presence of false positives; however, in this work we are interested in the shape and location of the composition distribution and can safely ignore the normalization constant needed for occurrence rate studies. For our purposes it is therefore sufficient to note that the probability of a given planet candidate being a false positive is roughly constant over our radius range ($\sim 5 - 10\%$; see §1.2.2), and so the presence of false positives are not expected to affect our results.

Second, we do not correct for transit probability, which is acceptable under the same conditions and for the same reason that we can ignore the false positive probability: our derivation of the composition distribution, which is a probability density function by definition, ignores the normalization factor central to occurrence rate studies. However, this is no longer the case if the planet radius is correlated with period, stellar type, or eccentricity, which would produce different transit probabilities for different radii. Of course, there are a number of reasons to expect that these correlations could exist due to the conditions under which these planets form and evolve; even our own incorporation of photoevaporation predicts a slight dependence between incident flux and composition (see Figure 5.10). Fortuitously, our results remain insensitive to the transit probability correction despite this: γ, which controls the rock-gas flux transition (Equation 5.7), is not correlated with the composition hyperparameters. Furthermore, it is a free parameter (which in our posterior samples varies between 2.2 and 3.0 to account for theoretical uncertainty in this threshold), and so much of the flux-composition

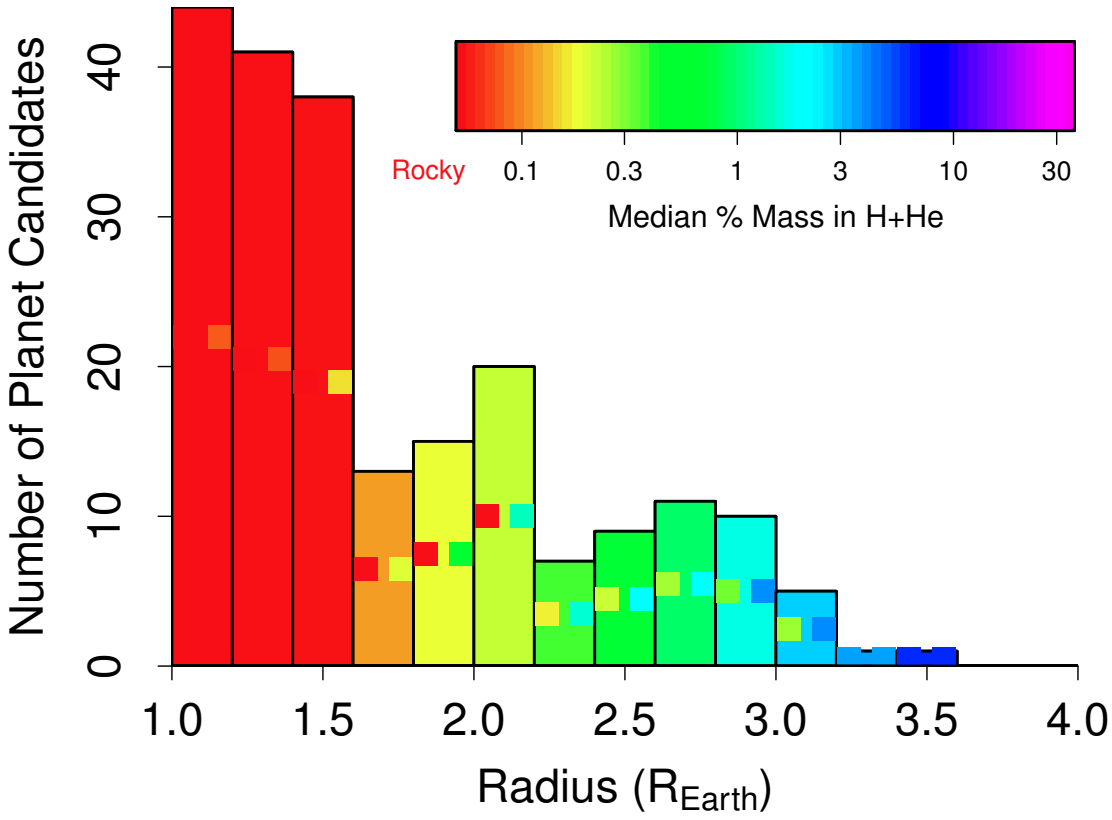

Figure 5.9: The radius distribution of our complete subsample, color coded according to the median composition in each bin and using the radii reported by *Kepler* (as opposed to the posterior radii in Table 5.1 and Figure 5.6) to facilitate comparison with Figure 5.1. The squares denote the full range of compositions within the bin: the lowest f_{env} in each bin corresponds to the left colored box, and the highest f_{env} the right. On average, interpreting radius as a proxy for composition is reasonable, although the large radius errors do allow some dispersion.

dependence present for individual planets gets washed out over the marginalized f_{env} posteriors in Table 5.1. Nevertheless, these concerns represent an interesting area for future work; correctly accounting for them requires modeling the underlying period, eccentricity, and host star radius distributions. Therefore, further development of this statistical model will require adding several additional layers of complexity to Figure

5.4.

Finally, we revisit the possibility that some of these planets may be water worlds. In §5.1.1 we motivated from an observational perspective why we assume all of these planets have a composition consisting of a rocky core and a hydrogen/helium envelope: for the first investigation of this population's composition distribution, it is natural to extend a two-component composition to the mid-range planetary sizes between the small, highly irradiated planets known to be rocky and the large low-mass sub-Neptunes which must have at least some H+He. Given these limits, it is difficult to motivate a population of sub-Neptunes that must all be characterized as water worlds. However, this does not mean that there cannot be a sub-population of water worlds, especially at periods longer than the planet candidates we consider in this work, and so this proposal is rich in possibilities for future work. Nevertheless, given the degeneracy (discussed in §5.1.1) between detailed compositions and measured mass and radius, an additional observable that can reliably distinguish between water-poor and water-rich bulk compositions will need to be measured and introduced to a statistical model like this one in order to get a quantitative handle on the extent of this possible sub-population.

5.5.2 Radius as a Proxy for Composition

A locus through (R_{pl}, f_{env}) space is immediately apparent in Figure 5.6, illustrating that "radius as a proxy for composition" is a reasonable interpretation to adopt for planets with $R_{pl} > 2$ $R_\oplus$, even with the current large, asymmetric errors on the planet radii. However, more variability is evident for smaller planets, especially in the $1.2 < R_{pl} < 1.8$ $R_\oplus$ range, where planets can either have rocky or gaseous compositions

(see §5.5.3 for a more detailed discussion). Given that the realistic radius errors included in this study does widen this locus, the following summary provides a reasonable rule-of-thumb when interpreting the composition of planets based on their radii: planets with $R_{pl} < 2$ R$_\oplus$ have $f_{env} < 1\%$, planets with $2 < R_{pl} < 3$ R$_\oplus$ have $f_{env} \sim 1\%$, and planets with $R_{pl} > 3$ R$_\oplus$ have $f_{env} \sim$ a few %.

Figure 5.9 further illustrates how the strong monotonic relationship between radius and composition can be extended to interpreting compositions from an observed radius histogram. We plot the radius distribution of our complete subsample (also shown in blue in Figure 5.1), but now color-code each bin according to the median composition of those planets, where a single value for composition, the mode of the f_{env} posterior, has been used for each planet. Taking radius as a proxy for composition would result in a monotonic increase in composition across the bins, which is exactly what we see. The picture complicates a bit when we consider the full range of compositions present in each bin, as illustrated by the colored boxes: the color of the left box corresponds to the lowest f_{env} in that bin, and the color of the right box corresponds to the highest f_{env}. The range within each bin illustrates the dispersion accommodated by the substantial errors on the planet radii. While the dispersion is currently non-negligible, it does not disrupt the average relationship between radius and composition.

5.5.3 The Rock-Gas Transition

Figure 5.6 also has implications for the expected transition between rocky and gaseous planets, assuming these planets do not have an appreciable mass fraction of water. In particular, we see that planets with $1.2 < R_{pl} < 1.8$ R$_\oplus$ can be either rocky

or gaseous, with f_{env} posteriors that span both compositions. This is consistent with the finding of Rogers (2015), which places the transition between rocky and gaseous planets at 1.5 $R_\oplus$ based on $\sim$ 50 *Kepler* confirmed planets with radial velocity mass measurements, primarily from Marcy et al. (2014). It is notable that internal structure models combined with the back-of-the-envelope parametrization of photoevaporation that we employ here (Equation 5.7) is able to reproduce this result within the context of these hierarchical MCMC simulations, given that we use no mass measurements to provide constraints as does Rogers (2015).

Our implementation of photoevaporation further predicts that there is some flux dependence to this transition, as seen in the color variation as a function of radius for the planets that could have either composition. Figure 5.10 more clearly illustrates this dependence: we plot the cumulative fraction of planets that are rocky in four flux bins, each containing 54 planets; the black line is the cumulative fraction for the entire sample. A planet is considered rocky if more than half of its f_{env} posterior occurs at 0, as is the case for the triangles in Figure 5.6. We see that the maximum radius for a rocky planet, denoted by the dotted vertical lines, increases slightly with increasing incident flux.

Rogers (2015) addressed this possibility by computing the marginal likelihood of the data under different hypothetical gas/rock transitions, including a sharp step function, a gradual linear relationship for the fraction of rocky planets as a function of radius, and a transition that depended on incident flux. With the existing large mass uncertainties, they find that the sharp transition is slightly favored over both

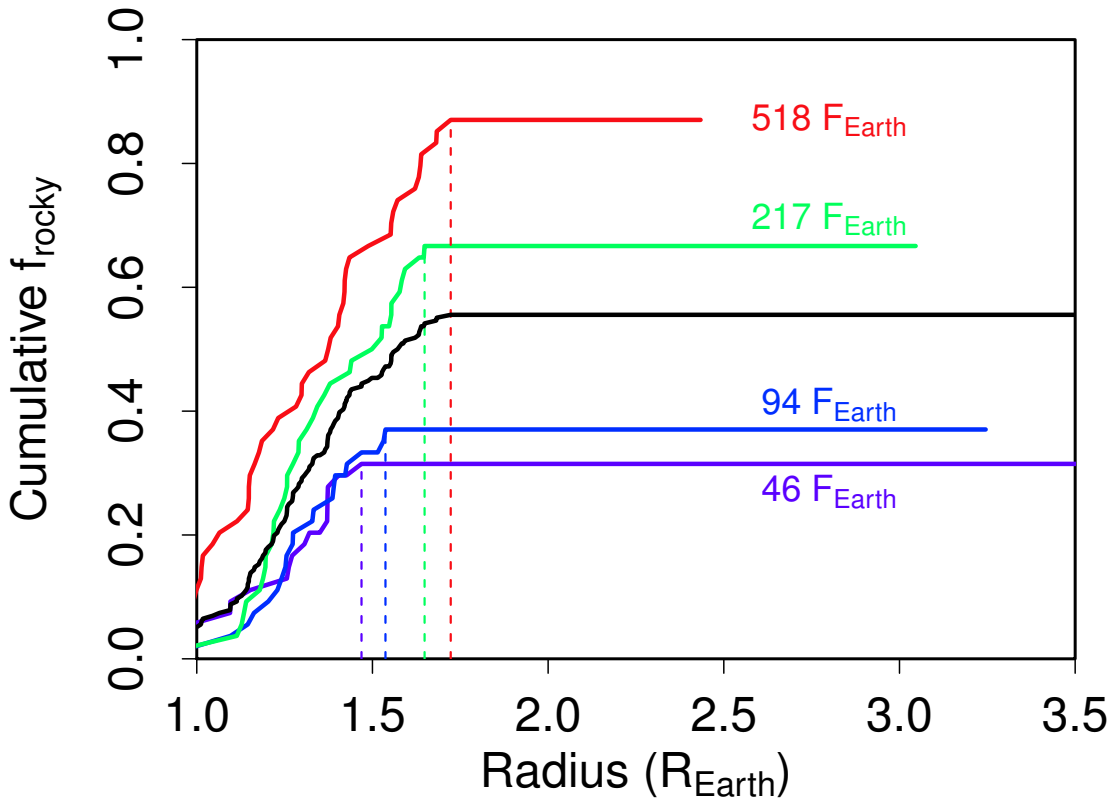

Figure 5.10: The cumulative fraction of planets that are rocky in four flux bins colored at the same scale as Figure 5.6 according to their labeled median flux values. Each colored bin contains 54 planets; the black line is the cumulative fraction for the entire sample. A planet is considered rocky if more than half of its f_{env} posterior occurs at 0. We see that the maximum radius for a rocky planet, denoted by the dotted vertical lines, increases slightly with increasing incident flux.

other options with a Bayes factor of ~ 2. We note that this Bayes factor is actually quite small for the purposes of inference, as one's prior belief in the realism of each of these transitions can still be a large factor in inferring which model best reflects what happens in nature. Furthermore, this factor can depend strongly on the choice of hyperprior, particularly when the prior is formally an improper distribution like the uniform distributions that were used. Therefore, we echo the author's caution that this

result does not mean the arguably more realistic transitions are ruled out, but rather that the currently large mass uncertainties do not allow one to distinguish between these possibilities.

Given the result of Rogers (2015), more precise mass measurements are needed before we can conclusively test the prediction that large rocky planets must have high incident stellar fluxes (for a complementary test of the rock-gas transition based on the differing tidal dissipation rates and resulting eccentricity distributions of rocky and gaseous planets, see Barnes (2014)). In particular, radial velocity follow-up of *Kepler* planet candidates can most effectively contribute to our understanding of photoevaporation by targeting $1.2 < R_{pl} < 1.8$ $R_\oplus$ planets at incident fluxes near this flux threshold. Two such planets are immediately identifiable in Figure 5.6, due to their high incident fluxes compared to the other similarly sized planets: KOI 171.01 (Kepler-116 b) and KOI 355.01, at 2.4 and 2.3 $R_\oplus$, and 470 and 440 $F_\oplus$, respectively. Because the mass loss flux threshold, and therefore the retention of the planet's envelope, is dependent on the core mass of the planet, we predict these planets must have fairly massive rocky cores, likely > 10 $M_\oplus$ (note that our simulations do not provide useful mass constraints for the rest of the smaller, less irradiated planets analyzed here, as the Lop14 internal structure models by themselves do not produce strong correlations between planet mass and radius — recall that this model feature allows us to derive compositions based mostly on radii in the first place). These planet candidates also happen to have fairly bright host stars, at a *Kepler* magnitude of 13.7 and 13.2, respectively, and so this prediction could in theory be tested with radial velocity measurements. Even if other observational

considerations cause mass measurements to be prohibitive for these specific planets, analogously large, highly irradiated super-Earths provide excellent leverage for testing theories of photoevaporation.

Regarding planets with massive cores, it is interesting to note that the most massive dense super-Earth found to date, Kepler-10c (Dumusque et al., 2014), would not in fact be rocky according to the models we use here. Based on its measured mass and radius ($\approx 17 \pm 2$ $M_\oplus$ and 2.35 $R_\oplus$), Kepler-10c should have a gaseous envelope fraction of $\sim 0.5\%$ (Lop14), or a relatively massive water steam envelope[9]. Rather than representing an extreme on the spectrum of possible super-Earth compositions, Kepler-10c instead exemplifies what we predict to be a fairly typical if somewhat massive sub-Neptune in terms of the envelope mass fraction it could possess.

5.5.4 No Deterministic Mass-Radius Relationship

Figure 5.11 illustrates what the sub-Neptune composition distribution that we find implies for the mass-radius relationship of these planets. Specifically, we generate a population of 10,000 planets using our "best fit" composition distribution ($\mu = 0.7\%; \sigma = 0.6$ dex) and a core mass distribution $\propto M^{-1}$, then randomly match these planets to the host stars and periods of the planets in our sample to apply the rock-gas transition flux threshold. The color corresponds to the generated planets' incident flux as in Figure 5.6, with the gradation at the low-mass end arising from our core mass-dependent prescription for photoevaporation. There is also a higher number density of

[9]For more examples of gaseous envelope fractions for planets with well measured masses, please see Table 7 of Lop14.

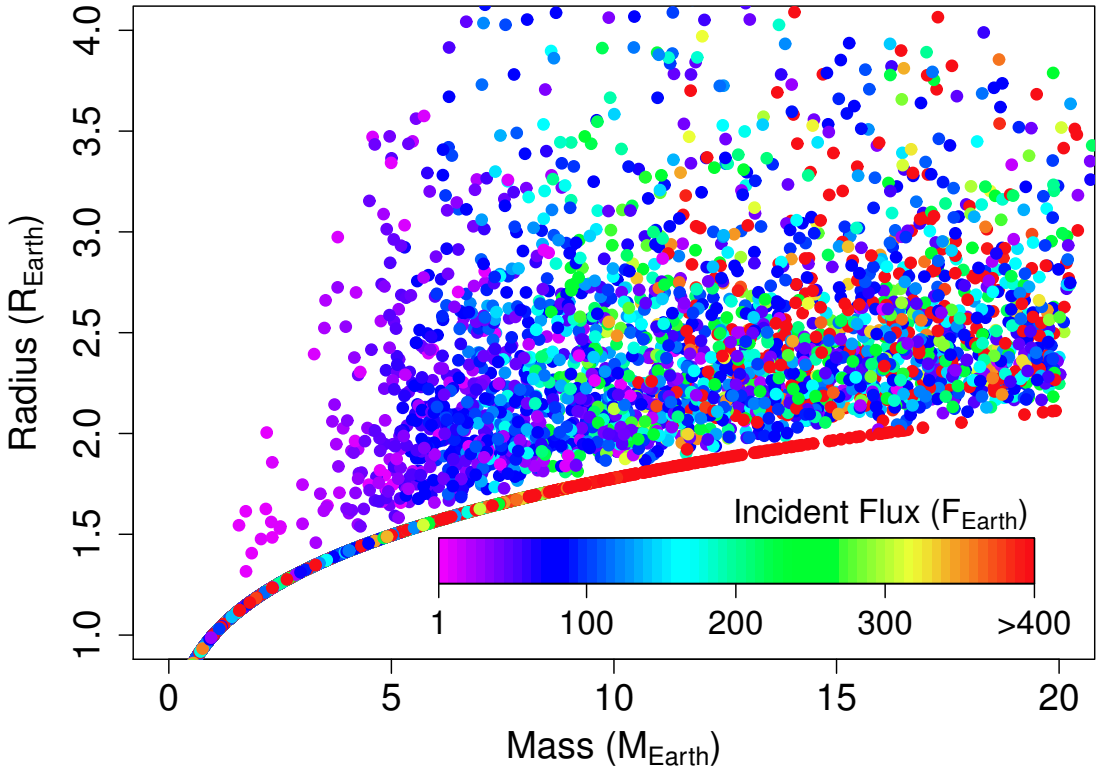

Figure 5.11: The masses and radii of a population of 10,000 planets generated from our "best fit" composition distribution (see Figure 5.5). Each point is colored according to its incident flux, at the same scale as Figure 5.6. Immediately we see that there is no clear mass-radius relationship for these sub-Neptune planets, although there are disallowed regions due to the maximum density of a rocky planet (at high masses and small radii) and to photoevaporation (at low masses and large radii). There is also a higher number density of planets between 2 and 2.5 $R_\oplus$; this is a direct result of our composition distribution peaking around $f_{env} \sim 1\%$.

planets between 2 and 2.5 $R_\oplus$; this is a direct result of our composition distribution peaking around $f_{env} \sim 1\%$.

Immediately we see that there is no clear one-to-one relationship, although there is a disallowed region at high masses and small radii, which is due to the max-

imum density of a rocky planet, and another at low masses and large radii, which is due to photoevaporation. While one could certainly fit a line to these points in mass and radius space, we argue that the more physically interesting variable at play is the composition. Because having a range of compositions dominates the spread in this plot (that is, the vertical extent of the mass-radius "relationship" is controlled by the distribution of compositions; see Figure 5.9 for another illustration of this), understanding planetary compositions in a population-wide sense requires robust statistical modeling that incorporates distributions rather than just mean relationships.

The lack of a deterministic mass-radius relationship for these sub-Neptune planets, which compose the majority of the planets that *Kepler* has detected, also has major implications for dynamical studies which require *Kepler* radii to be mapped to masses. Namely, such studies must adopt a probabilistic approach such as the one proposed in §3.5 to allow for a distribution of masses at a given radius. Without a way to incorporate the dispersion between mass and radius, the authors could be mislead by results that are seemingly more precise than they actually are. Similarly, theoretical studies could mistakenly rule out different parts of parameter space that may actually be allowed given the intrinsic uncertainty in the planet's mass based only on its radius.

5.5.5 Implications for Population Formation Models

Via our physically informed statistical modeling (§5.3.4) we have inferred the mean and variance of the present-day compositions of planets with 1 $R_\oplus < R_{pl} < 4$ $R_\oplus$, finding an average f_{env} of $\sim 1\%$ and standard deviation of ~ 0.5 dex, respectively (§5.4.1). As this result is derived directly from *Kepler* data, it offers a strong observa-

tional constraint for studies of planet formation which strive to characterize not only the average behavior of a few planets, but the range and distribution of various physically interesting variables across an entire planet population. Key diagnostics such as the range of compositions for these small planets can, for example, inform the degree of gas accretion during the planet formation process, and can therefore provide constraints on the relevant local protoplanetary disk parameters such as temperature and viscosity.

Of course, planetary evolution could also affect these planets' present-day compositions, and so the composition distribution we infer here has also encoded information about any of these processes which may have occurred. These quantitative constraints provide a first step in enabling comparisons between the effect that disk migration vs. multi-body interactions vs. in-situ formation could have on the amount of gas retained by super-Earths and sub-Neptunes, the most common kind of planets in our Galaxy. Much work remains to be done to disentangle these effects, and many other observational indicators such as spin-orbit misalignment and period ratios within multi-planet systems are being scrutinized. Nevertheless, with this analysis, planetary compositions can also enter into the conversation in a quantitative way.

5.6 Conclusions

In this paper we present the first quantitative distribution of sub-Neptune compositions. We find that, if these planets are composed of an Earth-like rocky core with a hydrogen and helium envelope, the "typical" sub-Neptune has $\sim 1\%$ of its mass in the gaseous envelope, while the population has a spread of ± 0.5 dex. We arrive at

this result by carefully choosing a subsample of *Kepler* planet candidates (§5.2) that is complete above 1.2 $R_\oplus$ (§5.2.1) and adopting a hierarchical Bayesian framework (§5.3) with a realistic yet relatively simple statistical model (§5.3.4) which incorporates the internal structure models of Lopez & Fortney (2014) and the stellar radius likelihoods derived by Huber et al. (2014). This approach simultaneously accounts for the lack of mass measurements and substantial radius measurement errors while describing the population-wide behavior with only four free parameters.

Our hierarchical Markov Chain Monte Carlo simulations (§5.3.5) result in posteriors on both the compositions of individual planets (§5.4.2) and on the composition distribution of the population (§5.4.1). Therefore, in addition to finding that the mean and standard deviation of the present-day compositions of planets with 1 $R_\oplus < R_{pl} < 4$ $R_\oplus$ is $\sim 1\%$ and ~ 0.5 dex, we can identify an honest the rule-of-thumb that relates radius to composition: planets with $R_{pl} < 2$ $R_\oplus$ have $f_{env} < 1\%$, planets with $2 < R_{pl} < 3$ $R_\oplus$ have $f_{env} \sim 1\%$, and planets with $R_{pl} > 3$ $R_\oplus$ have $f_{env} \sim$ a few %.

Finally, we discuss the implications that these results have for various issues related to the compositions of sub-Neptune planets. First, we verify that taking radius as a proxy for composition does hold up in the average sense even considering the large radius errors that exist for the majority of *Kepler* planet candidates (§5.5.2). We also address the rock-gas transition and discuss how carefully chosen and precise mass measurements could help test the theory of photoevaporation by elucidating a transition that is a function of incident flux (§5.5.3). In §5.5.4 we illustrate how this composition distribution means that there is no deterministic mass-to-radius relationship for

sub-Neptunes, and so dynamical studies must derive masses from *Kepler* radii probabilistically rather than with a simple one-to-one function. Finally, we discuss the rich opportunity these results offer for comparisons of planet formation studies with *Kepler*'s observed planetary candidates.

Chapter 6

Future Work

The thousands of irradiated super-Earth and sub-Neptune planets that *Kepler* has unearthed are extremely puzzling: they have no Solar System analogs in either size or period, yet they occur just as frequently as Earth-sized ones (Petigura et al., 2013) and could represent the default mode of planet formation. Their completely unexpected presence poses challenges to established paradigms of planetary accretion and migration, but even a zeroth-order understanding of why they are there and what they are like is elusive for several reasons:

- These planets are shaped by their radiative and gravitational environments after their natal protoplanetary disks have dispersed. Therefore, the late-stage evolution in these planets' radii and periods *must* be taken into account in order to tie currently observed properties to those that were a direct result of formation.

- Systematic effects, such as radius- and period-dependent detectability and incomplete performance of the automated detection software (Batalha et al., 2013),

permeate the *Kepler* sample. These effects manifest themselves in the observed distributions, and failure to account for them can mislead theoretical conclusions.

- There are substantial uncertainties on the planetary radii, masses, eccentricities, and other physical properties, often driven by the uncertainties in the host stars' properties. Errors can be as large as 100% (e.g. Huber et al., 2014; Marcy et al., 2014) and frustrate observers' efforts to draw statistically significant physical insight from their data.

With both TESS and CHEOPS slated for launch in 2017, the *Kepler* planet catalog will remain the preeminent dataset for comparison with theory for several years. Postdoctoral work to address these three issues is therefore very timely, and I am in a particularly good position to do it, given my joint expertise in the *Kepler* dataset and in the sophisticated statistical tools needed to address the last two issues above. I plan to use my NSF Postdoctoral Fellowship, which I am taking to Penn State, to start bridging the gap between observations and theory that is manifest in the first problem: how do super-Earths evolve after they form? In particular, I plan to investigate compositional evolution through photoevaporation and orbital evolution through Kozai-Lidov oscillations. By using the probabilistic framework developed for my thesis work, I will quantify, with realistic uncertainties, the fraction of planets which experienced these processes, therefore clarifying the possible end conditions to planet formation.

6.1 Framework for the Analysis of Planet Populations

My thesis work implements a framework that enables robust answers to population-wide questions about fundamental planet properties. In particular, Chapter 5 illustrates the need for a higher level of statistical sophistication than the current state of the art and uses hierarchical Bayesian modeling to infer the amount of gas that a typical sub-Neptune-sized planet possess. Chapter 3 applies this method to sub-Neptune-sized planets with measured masses to obtain a probabilistic mass-radius distribution.

In general, this hierarchical Bayesian framework establishes probabilistic relationships between data that are directly observed (i.e. transit depths) and quantities of theoretical interest (i.e. the fraction of a planet's mass that exists in a gaseous envelope, and the parameters which describe its distribution over the entire population). This necessarily involves other quantities that are either unobserved or have substantial observational uncertainties (i.e. planetary radii and masses). Due to its probabilistic construction, we can quantitatively derive population-wide distributions within this framework while correcting for both the systematic effects that permeate the *Kepler* planet candidate catalog (detailed in §1.2.2) and the substantial uncertainties on these planets' individual physical properties.

At the core of this work is a data-centric approach to the characterization of theoretically relevant planet properties. We endeavor to straddle the boundary between theory and observations: we use physically motivated arguments to synthesize theory into modules that are scalable to planet populations, and we incorporate observational uncertainties into the constraints on the theoretical parameters of interest (Chapter 5

provides a specific example of how this is done). I plan to continue to study *Kepler*'s sub-Neptune population to help answer the large number of fundamental science questions this unexpected population poses for planet formation and evolution, with the added benefit of its impressive sample size.

6.2 Compositional Evolution of Sub-Neptune-Sized Planets

With no Solar System analogs, the typical compositions of sub-Neptune-sized planets is a compelling question. Unfortunately, the difficulty in observing these extra-solar planets leads to relatively little information about them being available, especially compared to the Solar System planets; accordingly, we only have access to these planets bulk properties to provide insight into their structures and composition. Given the state of the observations, one common means of constraining these planets' interior compositions is to apply models of their internal structures to their measured masses and radii (e.g. Fortney et al., 2007; Rogers et al., 2011; Lopez & Fortney, 2014). This is usually accomplished for individual planets to gain insight into the range of possible compositions specific to that planet. However, the presence of intrinsic survey detection biases which cause planets with certain properties to be more easily detected and significant observational uncertainties for most sub-Neptune masses and radii make solving for the distribution of these compositions across an entire population of planets difficult.

In Chapter 5 I present the first study to rigorously account for these issues across an entire population; however, it treats only the present-day compositions of *Ke-*

pler's sub-Neptunes. Given their tight orbits around their host stars, photoevaporation of these planets' atmospheres are expected to be an important physical process for their past evolution. This is especially true in light of evidence for the hydrodynamic escape of hydrogen from HD 209458b (Vidal-Madjar et al., 2003), a hot Jupiter with a higher surface gravity, and thus a higher potential energy barrier, than these sub-Neptune-sized planets.

Because we aim to connect the current physical properties of these planets to the end conditions of their formation, we must assess how much mass these irradiated planets have lost since the dispersal of the protoplanetary disk. This investigation necessarily involves several inter-related projects, each targeting a different observational probe of this process. Systematic biases pervade each observational probe and must be corrected, which I will do with the population-wide statistical framework I discuss in §6.1. I will adopt the details of the framework, i.e. how the specific observables relate to the theoretical parameters of interest, to most appropriately address the problem at hand, as discussed below.

6.2.1 Dependence of Composition on Incident Flux

One of the most salient predictions of photoevaporation theory is that the composition of these sub-Neptunes should be correlated with the degree of irradiation they receive from their host star (e.g. Owen & Wu, 2013; Lopez & Fortney, 2013). A suggestive dearth in the *Kepler* period-radius distribution within $P \sim 3$ days for planets with $2\ \mathrm{R}_\oplus < R_{pl} < 4\ \mathrm{R}_\oplus$ has been noticed by these and other authors, and have been used to qualitatively guide appropriate choices for free parameters in the theory, such

as the efficiency of the thermal energy conversion needed to drive mass loss. However, these constraints are degenerate with the planets' masses, which are largely unknown, and do not incorporate uncertainties in the stellar parameters, which are substantial (Huber et al., 2014). Moreover, a recent study on the transition between rocky and gaseous planets finds no evidence for a transition that depends on incident flux, given the current mass measurements and a small sample size (Rogers, 2015).

A systematic search for variations in these planets' compositions as a function of incident flux is needed to clarify the picture; this will only be possible when the full sample of sub-Neptune-sized planets is used and when the uncertainties in planetary masses and stellar effective temperatures are taken into account through the framework described here. I plan to incorporate a bias- and completeness-corrected period distribution into my thesis work on sub-Neptune compositions to quantify the degree to which planetary radii (and by proxy, composition) depend on stellar irradiation. In doing so, I will relate the critical mass loss efficiency parameter to observed quantities and derive quantitative constraints on it based on the data. These results will also facilitate comparisons between photoevaporation studies employing the simplifying assumption of energy-limited mass loss, and more accurate yet computationally intensive simulations involving radiative transfer. Assessing the relative importance of such theoretical concerns requires a robust, quantitative application of theory to observations — the kind that only this framework can offer.

6.2.2 Population of Water-Dominated Planets

Studies investigating the compositions of sub-Neptune-sized planets must adopt a specific internal structure to establish quantitative constraints on the relative sizes of each layer within the planet. Unfortunately, the freedom in this choice makes the study of compositions a highly degenerate one (Valencia et al., 2007a). A rocky core plus a hydrogen envelope is a reasonable structure to assume for the sub-Neptune-sized population, given the highest and lowest densities that have been measured for these small planets. However, this does not mean that exoplanets which possess a substantial water layer do not exist, as both hydrogen-rock and water-rock compositions have been shown to be able to explain the observed masses and radii (Rogers & Seager, 2010a). With three free parameters and only two measured quantities, additional observational insight is needed to make concrete headway on this issue.

Optimally, observations of planetary atmospheres would reveal their dominant chemical species and offer valuable observational constraints on the extent of the water-dominated planet population. However, these measurements are currently pushing the limits of existing astronomical instrumentation and can only be performed in a limited number of cases, where the host star is bright, the planets surface gravity is low, and no high-altitude clouds or hazes are present. This last requirement is proving to be especially problematic, as clouds appear to envelop the majority of sub-Neptune-sized irradiated planets (e.g. Kreidberg et al., 2014). Even when features are observed in the wavelengths where one would expect water absorption, their interpretation is degenerate with a partly cloudy hydrogen-dominated atmosphere (Fraine et al., 2014).

In the absence of insight from studies of exoplanet atmospheres, photoevaporation theory offers a handle on this issue. Because hydrodynamic escape of water-dominated atmospheres is more difficult than it is for hydrogen-dominated atmospheres, highly irradiated planets that are measured to have fairly low densities ($\sim 1-3$ g/cm^3) are most likely such "water worlds" (e.g. Lopez et al., 2012). The above project feeds naturally into this investigation, as the lack of a radius-flux correlation can indicate the presence of a substantial water world population.

We can also gain insight into this problem by incorporating knowledge about where these planets had formed within their host stars' protoplanetary disks. This is possible because the available disk material at a planets' birth location sets its initial composition, and the disks own composition changes as a function of distance from the star (Aikawa & Herbst, 1999). As described in 6.3, we expect these planets to have experienced substantial orbital migration since their formation; if these planets originated from beyond the "snow line", i.e. the distance from the star at which water ice can condense (see, for example, Sasselov & Lecar 2000), then they could have water-dominated compositions. Therefore, the research outlined in the next section will also illuminate which planets could have substantial amounts of water, given their present-day orbital architectures.

With details contingent on the results of the other investigations detailed in this chapter, I plan to incorporate a population of water-dominated planets into our study of sub-Neptune-sized planet compositions. Our statistical framework can easily adjust to incorporate multiple groups of planets which possess qualitatively different

compositions. In particular, it will assign a probability that any given planet falls into each of these groups based on the likelihood that photoevaporation was inefficient and that the planet originated from beyond the snow line.

6.3 Orbital Evolution of Sub-Neptune-Sized Planets

It has been a challenge to explain the existence of the close-in planet populations revealed by various planet searches. As increasing numbers of hot Jupiters were discovered, the standard core accretion paradigm needed to be adjusted to incorporate significant orbital evolution (Lin et al., 1996). Two primary classes of theories emerged to fill this hole: disk migration (e.g. Goldreich & Tremaine, 1980; Ward, 1997; Ida & Lin, 2004; Alibert et al., 2005) and high eccentricity excitation mechanisms coupled with tidal dissipation and circularization of the planetary orbit (e.g. Rasio & Ford, 1996; Fabrycky & Tremaine, 2007; Wu & Lithwick, 2011). While these mechanisms have been applied to specific systems to show their feasibility (e.g. Holman et al., 1997; Lee & Peale, 2002; Wu & Murray, 2003), the picture is much less clear when the entire Hot Jupiter population is considered (e.g. Ford & Rasio, 2008), with varying levels of importance reported for various scenarios (Fabrycky & Winn, 2009; Morton & Johnson, 2011a; Naoz et al., 2012).

The plethora of sub-Neptune-sized planets that *Kepler* has unearthed provides an unrivaled opportunity to test the dominant mode of orbital evolution for a somewhat smaller yet prevalent planet population. Some groundwork has already been established. For example, Rein (2012) analyzed the period ratios of *Kepler* multiple-planet systems

and found that stochastic migration forces needed to be introduced to disk simulations to explain the observations. Later, Schlaufman (2014) used the frequency of long-period gas giant planets to argue that in-situ formation is not sufficient to explain the observed period distribution, assuming a disk with a smooth dust surface density profile.

6.3.1 Characterizing Distributions Relevant to Dynamics Studies

While these studies are a start, a significant amount of work is still needed to accurately characterize the current *Kepler* period, eccentricity, and multiplicity distributions themselves, as well as other observational indicators relevant to testing different migration mechanisms. There are a number of systematic effects present in the *Kepler* data that can easily mislead such comparisons to theory, including the heterogeneously selected target star sample (Batalha et al., 2010b), substantial uncertainties on stellar properties (e.g. Gaidos & Mann, 2013; Huber et al., 2014), lower detection efficiency at smaller radii and longer periods (e.g. Wolfgang & Laughlin, 2012; Howard et al., 2012), incomplete performance of the automated detection software (Batalha et al., 2013; Petigura et al., 2013), and the presence of false positives (Morton & Johnson, 2011b; Fressin et al., 2013). Without careful correction for each of these effects, the population distributions of these planets' physical properties will be biased, and conclusions with implications for theory are suspect.

To address this, Profs. Eric Ford (Penn State) and Darin Ragozzine (Florida Tech) have a funded program called SysSim to account for these effects and obtain the true period, eccentricity, and multiplicity distributions of *Kepler*'s planet candidates. As an NSF AA Postdoctoral Fellow at Penn State, I will contribute to this effort,

specifically by extending SysSim to include observational constraints on the presence of wide-binary companions. Afterwards, I will use the results as observational anchors for the dynamical calculations described below.

6.3.2 Probabilistic analysis of Kozai-Lidov oscillations

The hallmark of the high eccentricity migration family of theories mentioned in §6.3 is the presence of a stellar companion. A key distinguishing feature of one subclass of these theories, Kozai-Lidov cycles (Lidov, 1962; Kozai, 1962) with tidal friction (KCTF), is that the perturber which caused the migration should still be present within the system; furthermore, it should have a semimajor axis much greater than that of the planet, which ensures the long-term stability of this secular interaction. Given that the average distance to a *Kepler* target star is 1 kpc (Brown et al., 2011) and that the peak of the stellar binary period distribution lies at a separation of about 1000 AU (Duquennoy & Mayor, 1991), this prediction is directly and uniquely testable with adaptive optics (AO) follow-up observations of the *Kepler* field, such as that presented in Chapter 4.

While KCTF has largely been invoked to explain the presence of hot Jupiters and the observations of significant misalignment between these planets' orbits and the spin of their host stars (e.g. Fulton et al., 2013), this mechanism is independent of the planet's mass in the limit of a distant, massive perturbing companion (Fabrycky & Tremaine, 2007; Dawson & Chiang, 2014). Accordingly, star-induced KCTF could also be an important mechanism for *Kepler*'s Neptune-sized planets. KCTF furthermore requires the planet's eccentricity to achieve very high values at moderate planet-

star distances before experiencing tidal circularization to its present orbit, which could destabilize any other planets present at the distances that are observable by *Kepler*. Therefore, KCTF predicts that a higher proportion of *Kepler* single-planet systems should have stellar companions compared to multiple-planet systems, if it is in fact a dominant mechanism for orbital evolution. As it happens, *Kepler* has produced a significant population of Neptune-sized single planet candidates (Mullally et al., 2015), and we have preferentially observed these KOIs as part of our adaptive optics follow-up (Chapter 4). We therefore have a rich opportunity to test the generality of this evolution mechanism for *Kepler*'s single Neptunes.

While discovering bound stellar companions can provide constraints on the *a priori* likelihood that KCTF operated in this population, further characterization of these companions enables detailed dynamical simulations: the Kozai oscillation period depends on the mass, period, and eccentricity of the binary orbit, while the maximum eccentricity of the planet, and thus the final circularized semi-major axis, depends on the initial inclination between the binary and planetary orbits (Kiseleva et al., 1998). Comparing the output of these simulations with SysSim's true period, eccentricity, and multiplicity distributions (§6.3.1) within the analysis framework described here will provide both the probability that a given planet has undergone this type of orbital migration in its past, and the fraction of single Neptune systems which have experienced these cycles. Therefore, we can assess the overall importance of this mechanism for planetary orbital migration.

Of course, not all of these binary orbital quantities will be able to be tightly

constrained from our AO observations. Even worse, few, if any, of these stars will be confirmed as bound companions to begin with, given their large radial distances from Earth and the long time baselines of their potential orbits. Finally, AO non-detections do not rule out the present of fainter, closer companions to these host stars. These unfortunate realities are precisely why we need the statistical framework discussed in this chapter. In particular, HBM can:

- quantitatively and rigorously account for these individual systems' uncertainties, for example by incorporating probability distributions of inclination, eccentricity, and period given the observations;
- utilize additional information available for only a subset of the sample, such as relative color information that can constrain the physical distances between these targets and the detected sources, or forthcoming proper motion catalogs of *Kepler* targets that will provide evidence for common space motions of these stars;
- directly apply information about the population that is provided by previous work, such as the period and mass distributions of additional companions that have been found with radial velocity follow-up observations of Hot Jupiters (e.g. Knutson et al., 2014) and of *Kepler* planetary candidates (e.g. Marcy et al., 2014); and
- incorporate upper limits, as in the case where no stellar companions are detected, when sensitivity curves from the AO images can be used to place an upper limit on the mass of a potential perturber.

All told, the application of sophisticated, easily generalizable statistical frame-

works like hierarchical Bayesian modeling can lead to significant advances in the field of exoplanet astronomy. I am truly excited about the opportunities that will arise from more quantitative comparisons between theory and observations, and look forward to the improved understanding of planet formation that will no doubt result.

Bibliography

Adams, E. R., Ciardi, D. R., Dupree, A. K., et al. 2012, AJ, 144, 42

Adams, E. R., Dupree, A. K., Kulesa, C., & McCarthy, D. 2013, AJ, 146, 9

Aikawa, Y., & Herbst, E. 1999, A&A, 351, 233

Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, PASP, 125, 989

Alibert, Y., Mordasini, C., & Benz, W. 2011, A&A, 526, A63

Alibert, Y., Mordasini, C., Benz, W., & Winisdoerffer, C. 2005, A&A, 434, 343

Alonso, R., Brown, T. M., Torres, G., et al. 2004, ApJ, 613, L153

Bakos, G. Á., Lázár, J., Papp, I., Sári, P., & Green, E. M. 2002, PASP, 114, 974

Barnes, R. 2014, International Journal of Astrobiology, 14, 321

Basri, G., Walkowicz, L. M., Batalha, N., et al. 2011, AJ, 141, 20

Batalha, N. M., Rowe, J. F., Gilliland, R. L., et al. 2010a, ApJ, 713, L103

Batalha, N. M., Borucki, W. J., Koch, D. G., et al. 2010b, ApJ, 713, L109

Batalha, N. M., Borucki, W. J., Bryson, S. T., et al. 2011, ApJ, 729, 27

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, ApJS, 204, 24

Batygin, K. 2012, Nature, 491, 418

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Borucki, W. J., Koch, D. G., Basri, G., et al. 2011, ApJ, 736, 19

Brase, J. M., An, J., Avicola, K., et al. 1994, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 2201, Adaptive Optics in Astronomy, ed. M. A. Ealey & F. Merkle, 474–488

Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, AJ, 142, 112

Bryson, S. T., Tenenbaum, P., Jenkins, J. M., et al. 2010, ApJ, 713, L97

Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al. 2013, PASP, 125, 889

Burke, C. J. 2008, ApJ, 679, 1566

Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, ApJS, 210, 19

Butler, R. P., Marcy, G. W., Williams, E., et al. 1996, PASP, 108, 500

Campbell, B., Walker, G. A. H., & Yang, S. 1988, ApJ, 331, 902

Carter, J. A., Agol, E., Chaplin, W. J., et al. 2012, Science, 337, 556

Catanzarite, J., & Shao, M. 2011, ApJ, 738, 151

Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJ, 529, L45

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

Chatterjee, S., & Ford, E. B. 2015, ApJ, 803, 33

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

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, ApJS, 207, 35

Claret, A., & Bloemen, S. 2011, A&A, 529, A75

Coughlin, J. L., Thompson, S. E., Bryson, S. T., et al. 2014, AJ, 147, 119

Cumming, A., Butler, R. P., Marcy, G. W., et al. 2008, PASP, 120, 531

Dawson, R. I., & Chiang, E. 2014, Science, 346, 212

Dawson, R. I., & Johnson, J. A. 2012, ApJ, 756, 122

Demory, B.-O., Gillon, M., Deming, D., et al. 2011, A&A, 533, A114

Désert, J.-M., Charbonneau, D., Torres, G., et al. 2015, ApJ, 804, 59

Díaz, R. F., Almenara, J. M., Santerne, A., et al. 2014, MNRAS, 441, 983

Dong, S., & Zhu, Z. 2013, ApJ, 778, 53

Dressing, C. D., Adams, E. R., Dupree, A. K., Kulesa, C., & McCarthy, D. 2014, AJ, 148, 78

Dressing, C. D., & Charbonneau, D. 2013, ApJ, 767, 95

Dumusque, X., Bonomo, A. S., Haywood, R. D., et al. 2014, ApJ, 789, 154

Duquennoy, A., & Mayor, M. 1991, A&A, 248, 485

Everett, M. E., Barclay, T., Ciardi, D. R., et al. 2015, AJ, 149, 55

Fabrycky, D., & Tremaine, S. 2007, ApJ, 669, 1298

Fabrycky, D. C., & Winn, J. N. 2009, ApJ, 696, 1230

Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2014, ApJ, 790, 146

Fasano, G., & Franceschini, A. 1987, MNRAS, 225, 155

Figueira, P., Marmier, M., Boué, G., et al. 2012, A&A, 541, A139

Ford, E. B., & Rasio, F. A. 2008, ApJ, 686, 621

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64

Fortney, J. J., Marley, M. S., & Barnes, J. W. 2007, ApJ, 659, 1661

Fraine, J., Deming, D., Benneke, B., et al. 2014, Nature, 513, 526

Fressin, F., Torres, G., Charbonneau, D., et al. 2013, ApJ, 766, 81

Fulton, B. J., Howard, A. W., Winn, J. N., et al. 2013, ApJ, 772, 80

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

Gautier, III, T. N., Batalha, N. M., Borucki, W. J., et al. 2010, ArXiv e-prints, arXiv:1001.0352

Gavel, D., Kupke, R., Dillon, D., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9148, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 5

Gelman, A., & Rubin, D. 1992, Statistical Science, 7, 457

Goldreich, P., & Tremaine, S. 1980, ApJ, 241, 425

Guillot, T. 2010, A&A, 520, A27

Han, E., Wang, S. X., Wright, J. T., et al. 2014, PASP, 126, 827

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

—. 2013, ApJ, 775, 53

Ho, S., & Turner, E. L. 2011, ApJ, 739, 26

Hogg, D. W., Myers, A. D., & Bovy, J. 2010, ApJ, 725, 2166

Holman, M., Touma, J., & Tremaine, S. 1997, Nature, 386, 254

Horch, E. P., Howell, S. B., Everett, M. E., & Ciardi, D. R. 2014, ApJ, 795, 60

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

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

Howard, A. W., Sanchis-Ojeda, R., Marcy, G. W., et al. 2013, Nature, 503, 381

Howe, A. R., Burrows, A., & Verne, W. 2014, ApJ, 787, 173

Howell, S. B., Everett, M. E., Sherry, W., Horch, E., & Ciardi, D. R. 2011, AJ, 142, 19

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

Ida, S., & Lin, D. N. C. 2004, ApJ, 604, 388

—. 2010, ApJ, 719, 810

Jackson, B., Miller, N., Barnes, R., et al. 2010, MNRAS, 407, 910

Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al. 2010a, ApJ, 713, L120

—. 2010b, ApJ, 713, L87

Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010c, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 0

Jenkins, J. S., Twicken, J. D., Batalha, N. M., et al. 2015, AJ, arXiv:

Jin, S., Mordasini, C., Parmentier, V., et al. 2014, ApJ, 795, 65

Jontof-Hutter, D., Lissauer, J. J., Rowe, J. F., & Fabrycky, D. C. 2014, ApJ, 785, 15

Jurić, M., & Tremaine, S. 2008, ApJ, 686, 603

Kane, S. R., Ciardi, D. R., Gelino, D. M., & von Braun, K. 2012, MNRAS, 425, 757

Kelly, B. C. 2007, ApJ, 665, 1489

Kepler Mission Team. 2009, VizieR Online Data Catalog, 5133, 0

Kipping, D. M. 2014, MNRAS, 444, 2263

Kipping, D. M., Dunn, W. R., Jasinski, J. M., & Manthri, V. P. 2012, MNRAS, 421, 1166

Kiseleva, L. G., Eggleton, P. P., & Mikkola, S. 1998, MNRAS, 300, 292

Knutson, H. A., Fulton, B. J., Montet, B. T., et al. 2014, ApJ, 785, 126

Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, ApJ, 713, L79

Kornet, K., & Wolf, S. 2006, A&A, 454, 989

Kozai, Y. 1962, AJ, 67, 579

Kreidberg, L., Bean, J. L., Désert, J.-M., et al. 2014, Nature, 505, 69

Kurokawa, H., & Kaltenegger, L. 2013, MNRAS, 433, 3239

Lammer, H., Erkaev, N. V., Odert, P., et al. 2013, MNRAS, 430, 1247

Lammer, H., Selsis, F., Ribas, I., et al. 2003, ApJ, 598, L121

Latham, D. W., Stefanik, R. P., Mazeh, T., Mayor, M., & Burki, G. 1989, Nature, 339, 38

Latham, D. W., Rowe, J. F., Quinn, S. N., et al. 2011, ApJ, 732, L24

Law, N. M., Morton, T., Baranec, C., et al. 2014, ApJ, 791, 35

Lecavelier Des Etangs, A. 2007, A&A, 461, 1185

Lee, M. H., & Peale, S. J. 2002, ApJ, 567, 596

Léger, A., Selsis, F., Sotin, C., et al. 2004, Icarus, 169, 499

Léger, A., Rouan, D., Schneider, J., et al. 2009, A&A, 506, 287

Léger, A., Grasset, O., Fegley, B., et al. 2011, Icarus, 213, 1

Lidov, M. L. 1962, Planet. Space Sci., 9, 719

Lillo-Box, J., Barrado, D., & Bouy, H. 2012, A&A, 546, A10

—. 2014, A&A, 566, A103

Lin, D. N. C., Bodenheimer, P., & Richardson, D. C. 1996, Nature, 380, 606

Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al. 2011a, Nature, 470, 53

Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., et al. 2011b, ApJS, 197, 8

Lissauer, J. J., Marcy, G. W., Rowe, J. F., et al. 2012, ApJ, 750, 112

Lissauer, J. J., Marcy, G. W., Bryson, S. T., et al. 2014, ApJ, 784, 44

Lloyd, J. P., Liu, M. C., Macintosh, B. A., et al. 2000, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4008, Optical and IR Telescope Instrumentation and Detectors, ed. M. Iye & A. F. Moorwood, 814–821

Lopez, E. D., & Fortney, J. J. 2013, ApJ, 776, 2

—. 2014, ApJ, 792, 1

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

Loredo, T. J. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 371, Statistical Challenges in Modern Astronomy IV, ed. G. J. Babu & E. D. Feigelson, 121

Loredo, T. J. 2013, in Astrostatistical challenges for the new astronomy, Springer Ser. Astrostatistics (Springer, New York), 15–40

Lovis, C., Mayor, M., Bouchy, F., et al. 2009, in IAU Symposium, Vol. 253, IAU Symposium, ed. F. Pont, D. Sasselov, & M. J. Holman, 502–505

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. 2000, Statistics and Computing

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

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

Marley, M. S., Fortney, J. J., Hubickyj, O., Bodenheimer, P., & Lissauer, J. J. 2007, ApJ, 655, 541

Max, C. E., Olivier, S. S., Friedman, H. W., et al. 1997, Science, 277, 1649

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

Mayor, M., Udry, S., Lovis, C., et al. 2009, A&A, 493, 639

Mayor, M., Marmier, M., Lovis, C., et al. 2011, ArXiv e-prints, arXiv:1109.2497

Mazeh, T., Nachmani, G., Holczer, T., et al. 2013, ApJS, 208, 16

McGurk, R., Rockosi, C., Gavel, D., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9148, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 3

Moorhead, A. V., Ford, E. B., Morehead, R. C., et al. 2011, ApJS, 197, 1

Mordasini, C., Alibert, Y., Benz, W., & Naef, D. 2009, A&A, 501, 1161

Morton, T., Petigura, E., Johnson, J. A., et al. 2014, in American Astronomical Society Meeting Abstracts, Vol. 223, American Astronomical Society Meeting Abstracts #223, 206.06

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

Morton, T. D., & Johnson, J. A. 2011a, ApJ, 729, 138

—. 2011b, ApJ, 738, 170

Morton, T. D., & Swift, J. 2014, ApJ, 791, 10

Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, ApJS, 217, 31

Nagasawa, M., & Ida, S. 2011, ApJ, 742, 72

Naoz, S., Farr, W. M., & Rasio, F. A. 2012, ApJ, 754, L36

Nettelmann, N., Fortney, J. J., Kramm, U., & Redmer, R. 2011, ApJ, 733, 2

Nutzman, P., & Charbonneau, D. 2008, PASP, 120, 317

Olivier, S. S., An, J., Avicola, K., et al. 1994, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 2201, Adaptive Optics in Astronomy, ed. M. A. Ealey & F. Merkle, 1110–1120

Owen, J. E., & Jackson, A. P. 2012, MNRAS, 425, 2931

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

Pepe, F., Mayor, M., Delabre, B., et al. 2000, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4008, Optical and IR Telescope Instrumentation and Detectors, ed. M. Iye & A. F. Moorwood, 582–592

Pepe, F., Cameron, A. C., Latham, D. W., et al. 2013, Nature, 503, 377

Petigura, E. A., Marcy, G. W., & Howard, A. W. 2013, ApJ, 770, 69

Plummer, M. 2003, in Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 1, DSC 2003 Working Papers, ISSN:1609–395X

Pollacco, D. L., Skillen, I., Collier Cameron, A., et al. 2006, PASP, 118, 1407

Queloz, D., Casse, M., & Mayor, M. 1999, in Astronomical Society of the Pacific Conference Series, Vol. 185, IAU Colloq. 170: Precise Stellar Radial Velocities, ed. J. B. Hearnshaw & C. D. Scarfe, 13

Queloz, D., Bouchy, F., Moutou, C., et al. 2009, A&A, 506, 303

Quintana, E. V., Jenkins, J. M., Clarke, B. D., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1

Rasio, F. A., & Ford, E. B. 1996, Science, 274, 954

Rein, H. 2012, MNRAS, 427, L21

Ribas, I., Guinan, E. F., Güdel, M., & Audard, M. 2005, ApJ, 622, 680

Rivera, E. J., Lissauer, J. J., Butler, R. P., et al. 2005, ApJ, 634, 625

Rogers, L. A. 2015, ApJ, 801, 41

Rogers, L. A., Bodenheimer, P., Lissauer, J. J., & Seager, S. 2011, ApJ, 738, 59

Rogers, L. A., & Seager, S. 2010a, ApJ, 712, 974

—. 2010b, ApJ, 716, 1208

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

Rupprecht, G., Pepe, F., Mayor, M., et al. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5492, Ground-based Instrumentation for Astronomy, ed. A. F. M. Moorwood & M. Iye, 148–159

Santerne, A., Fressin, F., Díaz, R. F., et al. 2013, A&A, 557, A139

Santerne, A., Díaz, R. F., Moutou, C., et al. 2012, A&A, 545, A76

Santerne, A., Díaz, R. F., Almenara, J.-M., et al. 2015, ArXiv e-prints, arXiv:1505.02663

Sasselov, D. D., & Lecar, M. 2000, ApJ, 528, 995

Schaefer, L., & Fegley, B. 2009, ApJ, 703, L113

Schlaufman, K. C. 2010, ApJ, 719, 602

—. 2014, ApJ, 790, 91

Schlaufman, K. C., Lin, D. N. C., & Ida, S. 2009, ApJ, 691, 1322

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

Seager, S., Kuchner, M., Hier-Majumder, C. A., & Militzer, B. 2007, ApJ, 669, 1279

Seagroves, S., Harker, J., Laughlin, G., Lacy, J., & Castellano, T. 2003, PASP, 115, 1355

Ségransan, D., Mayor, M., Udry, S., et al. 2011, A&A, 535, A54

Sliski, D. H., & Kipping, D. M. 2014, ApJ, 788, 148

Srinath, S., McGurk, R., Rockosi, C., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9148, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2

Stetson, P. B. 1987, PASP, 99, 191

Strehl, K. 1902, Astronomische Nachrichten, 158, 89

Tabachnik, S., & Tremaine, S. 2002, MNRAS, 335, 151

Tenenbaum, P., Jenkins, J. M., Seader, S., et al. 2013, ApJS, 206, 5

—. 2014, ApJS, 211, 6

Torres, G., Fressin, F., Batalha, N. M., et al. 2011, ApJ, 727, 24

Tremaine, S., & Dong, S. 2012, AJ, 143, 94

Twicken, J. D., Chandrasekaran, H., Jenkins, J. M., et al. 2010a, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1

Twicken, J. D., Clarke, B. D., Bryson, S. T., et al. 2010b, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 23

Udalski, A., Paczynski, B., Zebrun, K., et al. 2002, Acta Astron., 52, 1

Udry, S. 2010, in The Theory and Observation of Exoplanets, Vol. 253, UC Santa Barbara KITP Programs, ed. A. Burrows, K. Menou, & D. J. Stevenson, 19

Udry, S., Mayor, M., Naef, D., et al. 2000, A&A, 356, 590

Valencia, D., Guillot, T., Parmentier, V., & Freedman, R. S. 2013, ApJ, 775, 10

Valencia, D., Ikoma, M., Guillot, T., & Nettelmann, N. 2010, A&A, 516, A20

Valencia, D., Sasselov, D. D., & O'Connell, R. J. 2007a, ApJ, 665, 1413

—. 2007b, ApJ, 656, 545

Vidal-Madjar, A., Lecavelier des Etangs, A., Désert, J.-M., et al. 2003, Nature, 422, 143

Wang, J., Fischer, D. A., Xie, J.-W., & Ciardi, D. R. 2014, ApJ, 791, 111

Ward, W. R. 1997, Icarus, 126, 261

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

Winn, J. N., Matthews, J. M., Dawson, R. I., et al. 2011, ApJ, 737, L18

Wittenmyer, R. A., Tinney, C. G., Butler, R. P., et al. 2011, ApJ, 738, 81

Wolfgang, A., & Laughlin, G. 2012, ApJ, 750, 148

Wolfgang, A., & Lopez, E. 2015, ApJ, arXiv:1409.2982, in press

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

Wright, J. T., Marcy, G. W., Howard, A. W., et al. 2012, ApJ, 753, 160

Wright, J. T., Fakhouri, O., Marcy, G. W., et al. 2011, PASP, 123, 412

Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 19

Wu, Y., & Lithwick, Y. 2011, ApJ, 735, 109

—. 2013, ApJ, 772, 74

Wu, Y., & Murray, N. 2003, ApJ, 589, 605

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