STATISTICAL STRATEGIES FOR CHARACTERIZING HABITABLE EXOPLANETS

by

David Alexander Bixel

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ASTRONOMY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN ASTRONOMY AND ASTROPHYSICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2021

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: David Alexander Bixel

titled: Statistical Strategies for Characterizing Habitable Exoplanets

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Daniel Apai   Date: May 13, 2021

Jared Males   Date: May 13, 2021

Regis Ferriere   Date: May 13, 2021

Glenn Schneider   Date: May 13, 2021

Mercedes Lopez-Morales   Date: May 13, 2021

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Daniel Apai   Date: May 13, 2021
Astronomy / Planetary Science

ACKNOWLEDGEMENTS

During the past five years I have had the privilege of contributing to one of the most profound and exciting endeavors in the history of science: the study of other worlds and the search for life beyond Earth. I am hopeful that my research will have a lasting impact on the field, and I would like to express my gratitude to the people who have enabled it.

First among these is Dániel Apai, whose best attribute as an advisor is his willingness to invite his students into exciting and ambitious projects at the ground floor. Dániel has helped me both to carve out my own niche in exoplanet research and to navigate my transition into a new career. I would similarly like to thank the members of my dissertation committee - Mercedes López-Morales, Glenn Schneider, Jared Males, and Regis Ferriere - for collaborating with me in research and guiding me toward completing my PhD.

Next, my friends and colleagues at Steward Observatory: Raga Pucha and Adam Sutherland, for navigating everything from lunch to prelims with me; Ben Rackham, for his patient support during my first year; Denise Gosé and Jeanne Davenport, for interfacing (and occasionally arguing) with the university on my behalf; and all of my office mates and board game friends.

Finally, my family and friends: my parents, who have always played an active role in my academic success; my sisters, grandparents, and my Uncle Robert; Jennie King, for suffering through undergrad physics with me; and Melissa Requist, for supporting me and keeping me sane throughout 2020.

DEDICATION

For my Grandpa, who reads all of my papers.

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF FIGURES – *Continued*

LIST OF FIGURES – *Continued*

LIST OF TABLES

ABSTRACT

The field of exoplanet astronomy is advancing toward the atmospheric characterization of Earth-like planets around nearby stars, with the ultimate goal of probing these exoplanets for evidence of life. Yet despite requiring the development of telescopes and measuring devices with unprecedented size and precision, this effort will yield only a fraction of the information content per planet afforded by robotic explorations of Solar System bodies. Given the diversity of terrestrial worlds within the Solar System, we can expect to face significant ambiguity in understanding the habitability and histories of potential exo-Earths, including potential "false positive" signatures of life. Given such limited data, how can we hope to determine which worlds are habitable or inhabited, or to understand more generally what factors make a planet habitable and give rise to life?

In this thesis, I propose solutions to this problem which draw upon the unique statistical advantage offered by their sheer numbers and the wide range of planetary properties they present. First, I demonstrate how existing knowledge about the frequency and sizes of terrestrial planets can be used to probabilistically constrain the composition of Proxima Centauri b - a nearby planet for which we have few direct measurements - and find that it is quite likely to be a terrestrial planet. Next, I demonstrate how a similar approach could be combined with future direct imaging observations of nearby stellar systems to determine which of their planets are most likely to be potentially habitable. This optimized target selection strategy could save weeks to months of follow-up observing time on a flagship-class space telescope. Following that, I present my analysis of the transit spectrum of WASP-4b, in which I approach the critical unresolved issue of stellar contamination which could limit the usefulness of this technique for studying habitable exoplanets in the future. My Bayesian evidence-based approach presents a possible solution for future analyses. In the final chapters, I focus on developing testable statistical hypotheses for future surveys of habitable exoplanets which would shed light on how these worlds

form and evolve. I begin by proposing that a correlation might exist between the ages of habitable planets and the fraction which have oxygen, which I dub the "age-oxygen correlation". A successful test of this hypothesis would demonstrate that other inhabited planets evolve in similar ways to Earth, and would suggest that atmospheric O_2 can be interpreted as robust evidence of life. Next, I expand in this direction by developing a general framework for evaluating the potential of next-generation space telescopes to test statistical hypotheses such as these. I apply the framework to demonstrate the requirements for an observatory to detect the existence of the habitable zone and constrain its boundaries, as well as to measure the timescale of atmospheric evolution on Earth-like planets. Finally, I condense my results into key recommendations for future efforts to study habitable exoplanets and search for life beyond the solar system.

CHAPTER 1

Introduction

1.1 Habitable worlds in the solar system and beyond

To date, Earth remains the only planet known to have ever sustained life. As such, it has traditionally been used as the model for a habitable world in the search for life elsewhere. Though a multitude of factors may affect the ability of life to flourish, the presence of liquid water has been highlighted as the most important criterion for habitability, as its unique properties as a chemical solvent have made it a necessity for all life on Earth.

Based on this criterion, the solar system may have hosted a multitude of habitable worlds throughout its history. Both Mars and Venus may have been habitable in the past, and both were rendered uninhabitable on their surface for different reasons. On Mars, geological features show increasingly clear evidence of flowing water in its early history (Malin and Edgett, 2003), and perhaps an ocean in the northern hemisphere (Fawdon et al., 2018). Due both to its lack of a magnetic field and its smaller size and surface gravity, Mars lost most of its atmosphere to stripping by the solar wind, and with it the potential for life on its surface (Jakosky et al., 2015, 2017). Recent modeling suggests Venus may have been habitable as recently as ~ 1 billion years ago (Way et al., 2016), but it was eventually pushed into a runaway phase, during which its water evaporated, and after which a thick CO_2 atmosphere formed from the carbon in its mantle. These two examples provide clues into the key processes affecting planetary habitability and the size and insolation limits beyond which habitability breaks down. Modern-day potentially habitable worlds in the solar system include Titan (which has liquid methane rather than water on its surface) and a number of bodies that might host sub-surface water oceans, including the moons Europa, Enceladus, and Ganymede. The discovery of forms of life with a unique origin to Earth's in any of these environments would suggest that life-bearing worlds should exist throughout the Milky Way.

However, due to the finite number of planets and moons it contains, even an exhaustive exploration of the solar system would still leave us with an incomplete understanding of the factors that make planets habitable and the conditions under which life arises. Today, the solar system hosts just one planet with a temperate climate and large bodies of liquid surface water. All of its planets are approximately the same age, and information about their pasts can only be inferred through geological evidence and evolutionary models. The Sun is a warm G dwarf star, whereas most stars are cooler M dwarfs, and there likely exist many types of terrestrial planets not represented in the solar system at all. Finally, Earth may turn out to be the only inhabited world in the solar system, in which case the search for extraterrestrial life would need to extend beyond our Sun.

By contrast, the number of planets discovered around other stars (extra-solar planets or exoplanets) currently numbers over 4,000[1] and continues to grow. This census includes a number of potentially habitable worlds, such as systems of multiple temperate Earth-sized planets (Gillon et al., 2017), tidally-locked planets with permanent day and night sides (Turbet et al., 2016), and planets around highly active flare stars (Howard et al., 2018). From these discoveries we can infer that at least millions of potentially habitable worlds exist within the Milky Way, and efforts are already underway to discover the nearest ones and probe them for signs of life. Due to the immense distances between the stars, our knowledge about any one of these exoplanets will be limited - nevertheless, their sheer number and diversity offer the only path toward a complete description of habitability across the full range of planetary parameters.

This thesis focuses on strategies for leveraging the powerful statistical advantage offered by exoplanets toward the specific goal of characterizing habitable worlds. In the work presented here, I use statistical methods to better understand the properties of recently-discovered planets, to propose more efficient methods for identifying habitable worlds using next-generation observatories, and to develop new hypotheses for how habitable planets evolve and determine what types of new observatories would be required to test them. In this chapter, I will summarize current knowledge about habitable exoplanets and review our prospects for characterizing them in the future. Following a brief review

---

[1]Per the NASA Exoplanet Archive.

of Earth's history and its relevance to the search for Earth-like planets in Section 1.2, I will summarize some recent findings in the field of exoplanet statistics in Section 1.3. In Section 1.4, I will discuss the most promising techniques for characterizing habitable exoplanets and concepts for next-generation observatories that would leverage them. Finally, in Section 1.5, I will describe the outline for the rest of the thesis.

1.2   Co-evolution of Earth and its biosphere[2]

Earth as we know it today, with its temperate climate and oxygen-rich atmosphere, is only one realization of a constantly evolving world. Up until 500 million years ago, Earth would have been uninhabitable to humans as there was far too little oxygen to sustain our respiration. Due to the steady increase in the Sun's luminosity, Earth will likely lose its water within 1–2 billion years, rendering it permanently uninhabitable (Caldeira and Kasting, 1992). This means that for the majority of the Sun's lifetime, Earth was or will be a markedly different planet than it is at present, and this fact has important implications for the search for Earth-like exoplanets.

Following the tumultuous events of Earth's formation and the moon-forming impact, life began to flourish during the late Hadean or early Archean eras under a predominantly nitrogen and carbon dioxide atmosphere. At first, life converted sunlight into chemical energy through anoxygenic photosynthesis, a primitive form of photosynthesis that does not produce O_2. Before long, however, "whiffs" of oxygen would appear in localized regions, hinting at the evolutionary development of oxygenic photosynthesis, the mechanism by which most energy enters the ecosystem today (Anbar et al., 2007). This oxygen was at first unable to spread globally, as it was quickly consumed through oxidation reactions with rocks on the surface and reduced gasses in the atmosphere.

The transition into the Proterozoic era was marked by the most dramatic change in the history of Earth's atmosphere - one that was closely linked to the metabolic processes of life (Figure 1.1). Over a brief time period, the oxygen content of the atmosphere spiked from effectively zero to approximately $0.001 - 1\%$ (by volume) during the "Great

---

[2]This section draws from Lyons et al. (2014), to which I refer the reader for a more in-depth review.
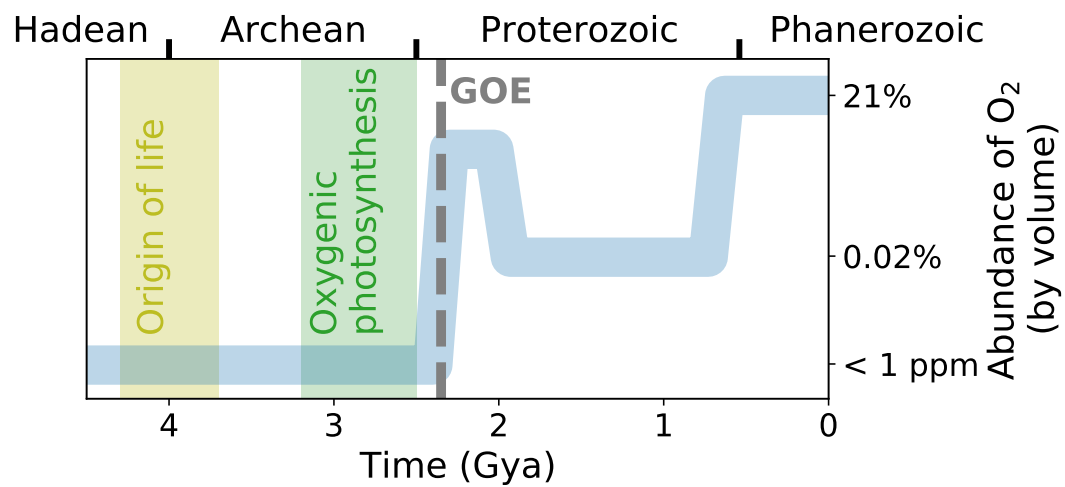
Figure 1.1: The history of oxygen in Earth's atmosphere since 4.5 billion years ago (Gya) with major eras and events labeled, adapted from Lyons et al. (2014). Values before ~ 0.5 Gya are approximations or upper limits. Life most likely arose in the late Hadean or early Archean era in an atmosphere rich in N_2 and CO_2. Oxygen, produced through photosynthesis, appeared in local regions during the late Archean, but did not accumulate globally until the Great Oxygenation Event at the beginning of the Proterozoic era. Another rapid oxygenation event heralded the beginning of the Phanerozoic era and led to the development of complex forms of life. Could similar evolutionary histories occur on inhabited exo-Earths?

Oxygenation Event" (GOE). What exactly caused the GOE remains under debate because, as mentioned earlier, oxygen-producing life had likely already existed for several millions of years. The oxygen content of the atmosphere is determined by the balance between its production by photosynthetic life and its destruction by the geological sinks described above. Therefore, it is necessary that either photosynthesis rapidly expanded and/or those sinks were dramatically reduced. For example, it has been suggested that the escape of hydrogen from the upper atmosphere led to a long-term decline in Earth's inventory of hydrogen-rich reduced gasses, until it was insufficient to counter biological oxygen buildup (Catling et al., 2001). Finally, the transition into the Phanerozoic era corresponded with another increase in atmospheric oxygen, this time to its modern value of $\sim 20\%$. The appearance of complex life, including animals, would soon follow. Today, Earth's oxygen content is largely determined by the balance between its production through photosynthesis and its consumption through respiration - another biological process. From this synopsis, it is clear that Earth's atmospheric evolution has been directly tied to the presence of life.

Earth's dramatic evolutionary history both complicates and motivates the search for Earth-like exoplanets. On the one hand, given that known exoplanets range from millions to billions of years in age, we may be unlikely to find many that resemble Earth in its modern state. In particular, O_2 (and its byproduct O_3, which is produced through photochemistry in the upper atmosphere) is viewed favorably as a potential biosignature due to its strong biological sources on Earth - but based on Earth's own history, it may be the case that even planets teeming with life have little to no oxygen in their atmospheres (Reinhard et al., 2017). For these planets, alternative biosignatures indicating strong chemical disequilibrium must be explored, such as the simultaneous presence of CH_4 and CO_2 (Krissansen-Totton et al., 2018). On the other hand, the study of Earth's history offers us multiple examples of what habitable exoplanets might look like and the ways in which life can affect their atmospheres over time. For example, in Chapter 5, I show how a statistical analysis of the atmospheres of other habitable worlds could reveal whether or not they, too, have co-evolved with life toward greater oxygen content. This would require the detection of Earth-like planets of varying ages with both oxygen-free
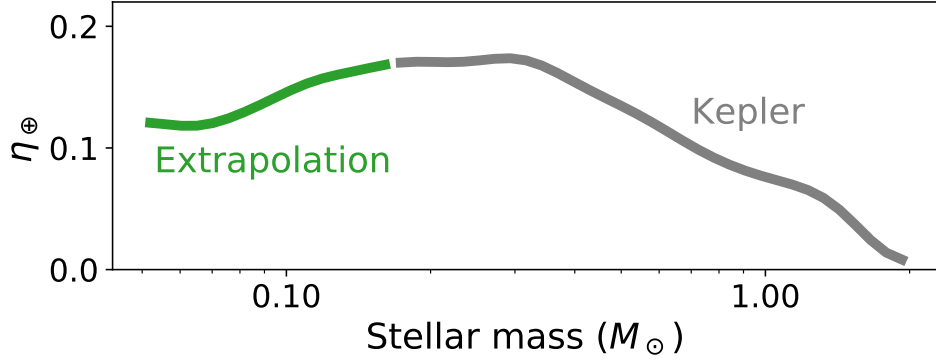
Figure 1.2: An estimate for the average number of potentially habitable planets per star ($\eta_\oplus$) as a function of stellar mass. A potentially habitable planet is defined as having a size between $\sim 0.8 - 1.4 R_\oplus$ and an orbit within the habitable zone (Kopparapu et al., 2014). The grey line is based on *Kepler* data and accounts for the bias due to photoevaporation highlighted by Pascucci et al. (2019), while the green line is a conservative extrapolation for ultra-cool dwarf stars. These values were produced using the `Bioverse` code presented in Chapter 6.

and oxygen-rich atmospheres. On balance, the expected diversity of evolutionary states among habitable planets could greatly enhance the study of habitable worlds, but only if enough such planets are observed to properly characterize that diversity.

1.3 How common are potentially habitable planets?

The field of exoplanet statistics seeks to understand the frequency of different types of planets and how their measurable properties relate to one another. This effort benefited immensely from the discovery of thousands of exoplanets by the *Kepler* Space Telescope (Borucki et al., 2010) as well as multiple ground-based surveys to characterize nearby exoplanets and refine measurements of *Kepler* host star properties. Statistical studies based on these datasets have granted us a more nuanced understanding of which planetary and stellar properties affect habitability, and to derive the first empirical estimates for the frequency of potentially habitable planets orbiting nearby stars.

1.3.1 *Kepler* occurrence rates and $\eta_\oplus$

The *Kepler* Space Telescope was launched with the goal of studying the frequency and distribution of extrasolar planets, with a particular emphasis on determining the number of Earth-sized planets in the habitable zone of a typical Sun-like star - with the "habitable zone", discussed below, being the range of distances from a star over which an Earth-like planet could maintain liquid water on its surface. To do so, it continuously monitored hundreds of thousands of stars within a single patch of the sky looking for transit events, which occur when a planet orbits in front of its host star as viewed from Earth. During the transit event, the planet temporarily blocks a small fraction of the star's light, which can be measured even though the planet itself cannot be resolved by the telescope. These events occur once for each orbit the planet completes, and the fraction of light blocked depends on the surface area of the planet relative to its star. As a result, *Kepler* was not only able to detect thousands of exoplanets through the transit method, but also to measure their periods and radii with high precision.

The primary results of the *Kepler* survey can be summarized by occurrence rates, which quantify the average number of planets per star as a function of orbital period and size. Occurrence rates are not trivial to derive from the *Kepler* catalog of detected planets, as *Kepler* was biased toward the detection of short-period planets (which are more likely to transit their stars and do so more frequently) and large planets (which block a larger fraction of the stellar light). Any realistic analysis must quantify and correct for these biases. A useful summary statistic of *Kepler* occurrence rates is $\eta_\oplus$ (pronounced "eta Earth"), defined as the average number of planets per Sun-like star whose orbital periods lie within the habitable zone and whose sizes are similar enough to Earth that they might be habitable. The precise definition of $\eta_\oplus$ depends on assumptions made about the habitable zone and the relationship between planet size and habitability, discussed below. However, *Kepler* detected few planets around Sun-like stars whose orbital periods place them within the habitable zone, and all of them are much larger than Earth. In order to estimate $\eta_\oplus$, one must extrapolate from detections of planets with shorter periods and/or larger sizes, and differences in extrapolation methods have led to varying estimates of $\eta_\oplus$.

The true value of $\eta_\oplus$ is of high significance to NASA and other organizations seeking to construct space telescopes capable of characterizing habitable planets, as it determines the expected number of potential targets as a function of telescope design parameters. Therefore, NASA's Exoplanet Exploration Program Analysis Group chartered a Study Analysis Group (SAG 13) to consolidate these disparate results into a single set of estimates for $\eta_\oplus$[3]. The "consensus" value that emerged was $\eta_\oplus \approx 0.24$, suggesting one potentially Earth-like planet for every four Sun-like stars. However, evidence has since arisen that most existing estimates of $\eta_\oplus$, including those consolidated by SAG 13, do not sufficiently account for the effect of atmospheric loss on the period-radius distribution of *Kepler* planets (Pascucci et al., 2019). Specifically, many short-period planets with large atmospheres will end up losing them due to photoevaporation, leaving behind their small, rocky cores. Extrapolations from this short period regime will therefore overestimate the number of small, rocky planets within the habitable zone, where the stellar flux is lower and more planets are able to maintain large atmospheres. On the other hand, a more recent analysis that excluded planets receiving high amounts of flux (thus minimizing bias due to photoevaporation) suggests values as high as $\eta_\oplus \sim 1$ - albeit with larger uncertainties due to a limited sample size (Bryson et al., 2021). Despite the overall success of *Kepler*, the true value of $\eta_\oplus$ remains only approximately constrained.

Other studies of *Kepler* occurrence rates focus on their dependence on stellar mass. *Kepler* was biased toward the detection of planets around brighter stars such as the Sun, and found a comparatively lower number of planets around low-mass M dwarfs. However, Mulders et al. (2015b) showed that the average number of terrestrial planets orbiting these low-mass stars is $\sim 3.5\times$ as high as for Sun-like stars, suggesting a similar enhancement to $\eta_\oplus$ for these systems. In Figure 1.2, I combine results from their analysis, the SAG 13 metastudy, and Pascucci et al. (2019) to derive estimates for the value of $\eta_\oplus$ as a function of stellar mass. The relative ease of detecting potentially habitable planets orbiting low-mass stars, combined with their higher predicted numbers, makes them attractive targets for atmospheric characterization in the near future.

---

[3]See this URL and Kopparapu et al. (2018).
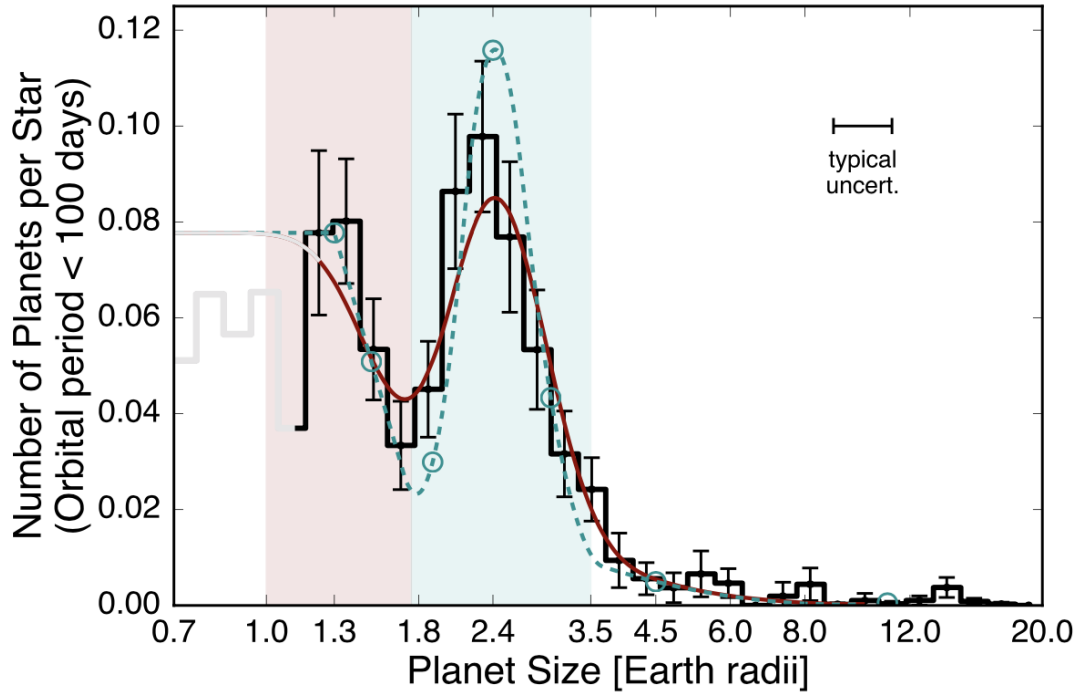
Figure 1.3: The size distribution of *Kepler* exoplanets with periods shorter than 100 days, from Fulton et al. (2017). The bimodal distribution shows two broad classes of small exoplanets: smaller rocky planets with compositions similar to Earth (red), and larger ice giants closer in composition to Neptune (blue). This suggests an upper limit on the size of habitable planets between $1.4 - 1.7\ R_\oplus$.

1.3.2 Habitability and planet size

By using the transit method, *Kepler* was able to measure its planets' sizes, but not their masses, densities, or atmospheric properties. As such, any estimate of $\eta_\oplus$ depends on some assumptions about the minimum and maximum radii which a habitable planet could have.

A common definition of the minimum habitable radius is that below which the planet is likely unable to withstand atmospheric stripping by its host star. With insufficient atmospheric pressure, any liquid water on the surface would immediately evaporate. One example of such a planet can be found in Mars, which almost certainly possessed a more substantial atmosphere along with lakes and rivers of liquid water during its early history, but whose atmosphere has since been stripped by the solar wind. Zahnle and Catling

(2017) pointed out that the planets and moons in the solar system that have atmospheres can be neatly divided from those without when plotting their insolations (I) versus escape velocities (v_{esc}). The dividing line can be expressed as:

$$I \propto v_{\mathrm{esc}}^4 \tag{1.1}$$

Furthermore, exoplanets with measured radii and masses are found to crowd against this line, suggesting it might be applicable beyond the solar system. The importance of insolation and escape velocity demonstrates that atmospheric escape is the key factor determining which planets have atmospheres, rather than the conditions of their formation. Applying this rule to rocky planets and assuming a fixed density, one can derive a minimum radius below which planets are unlikely to maintain atmospheres:

$$R < (0.8\,R_\oplus)a_{\mathrm{eff}}^{-1/2} \tag{1.2}$$

where $a_{\mathrm{eff}} = (a/1\mathrm{AU})(L_*/L_\odot)^{-1/2}$ is the semi-major axis (a, in AU) normalized by the luminosity of the host star (L_*).

One possible upper limit on a habitable planet's size can be derived from the subset of transiting planets whose masses can be constrained through radial velocity or transit timing variation measurements. Combining their measured masses and radii yields constraints on their densities, and therefore their likely composition. Weiss and Marcy (2014) and Rogers (2015) noted that planets larger than approximately 1.5 $R_\oplus$ begin to take on low bulk densities more consistent with Neptune or Uranus than Earth. More recently, Fulton et al. (2017) found evidence that two distinct populations of planets existed in the *Kepler* sample - one with smaller sizes and the other with larger sizes, shown in Figure 1.3. The emerging consensus from these results is that a dichotomy exists between smaller rocky planets and larger ice giants, with the dividing radius placed somewhere between $1.4 - 1.7\,R_\oplus$. Planets larger than this have sufficient gravity to maintain thick, H/He dominated atmospheres, while smaller planets will lose these thick envelopes but retain more massive species such as N_2. The larger planets, which likely resemble Neptune, would not be hospitable to Earth-like life.

1.3.3 The habitable zone

While we have some data to suggest how habitability varies with planet size, we have very little to inform us as to the location of the Sun's habitable zone (except that, by definition, it must include Earth). As a result, estimates of the habitable zone's boundaries are typically based on modeling the energy balance of an Earth-like planet at varying distances from the Sun. The most commonly cited estimate ($\sim 0.95 - 1.7$ AU, Kopparapu et al., 2013, 2014) is based on a model of the carbon cycle that has been successfully used to describe Earth's climate history. In this model, atmospheric CO_2 is dissolved into rainwater, which precipitates it onto to the surface where it undergoes weathering reactions with rocks. The carbonate products of these reactions are then carried by streams and rivers into the ocean, where they are ultimately deposited onto the seafloor. Over long timescales, the carbon will be subducted into the mantle, where CO_2 is reforged and returned to the atmosphere through volcanic outgassing. Crucially, both the amount of rainfall and the rate of weathering reactions increase with temperature, leading to a stabilizing feedback mechanism: if Earth warms up, the deposition rate of atmospheric CO_2 increases, leading to a weaker greenhouse effect which acts against the warming trend. The opposite is true as well: if Earth cools down, the carbon sinks become less efficient, allowing an increase in atmospheric CO_2 through volcanic outgassing and a stronger greenhouse effect to counteract the cooling trend (Walker et al., 1981). This stabilizing feedback, in principle, would allow an Earth-like planet to maintain habitable surface temperatures much farther from the Sun than 1 AU, leading to a broad habitable zone (Kasting et al., 1993).

The boundaries of the habitable zone are the limits beyond which this feedback breaks down. Planets much closer than ~ 0.95 AU are predicted to undergo a runaway greenhouse effect, causing the oceans to rapidly evaporate. Once all of the planet's water is evaporated, solar photons will split H_2O, and the lone H atoms will be lost to space. Without hydrogen, the planet's water cannot be reconstituted, thus it has undergone permanent water loss. Meanwhile, planets farther than ~ 1.7 AU become so cold in their upper atmospheres than even CO_2 cannot maintain a gaseous form. The CO_2 condenses

into clouds, which provide no greenhouse effect while also increasing the planet's reflectivity, thus reducing its temperature further. Eventually, all of the CO_2 condenses and the planet freezes over. No such frozen Earth-like planets exist in the solar system, though they might exist around other stars (Kasting et al., 1993).

The predicted range of $0.95 - 1.7$ AU is valid for a Sun-like star, but the habitable zone is also a function of stellar mass. Lower-mass stars are far less luminous, so the habitable zone is much closer in (approximately $\sim 0.1 - 0.2$ AU for a typical red dwarf; Kopparapu et al., 2014). The habitable zone is additionally dependent on the planet's reflectivity, volcanic outgassing rates, and other factors affecting its energy balance that are not trivial to measure. Critically, all habitable zone estimates are model-derived, as we lack the means to empirically test them. Such a test is, however, potentially within the scope of future observatories designed to characterize habitable worlds. In Chapter 6, I suggest a simple test to detect the habitable zone and constrain its boundaries that involves studying how the presence of water vapor in exoplanet atmospheres varies with their distances from their stars. Others have proposed to investigate whether CO_2 concentration truly does increase for planets farther from their stars as climate models predict (Bean et al., 2017; Checlair et al., 2019; Lehmer et al., 2020). In either case, the existence and boundaries of the habitable zone cannot be inferred through observations of any individual planet, but only by analyzing a statistically relevant sample.

Combining these lines of evidence, we come to a heuristic definition for a *potentially* habitable planet: that it resides within the circumstellar habitable zone defined by Kopparapu et al. (2014) and has a radius between $0.8a_{\text{eff}}^{-1/2} - 1.4 \, R_{\oplus}$ (i.e. a conservative interpretation of the maximum size of a rocky planet). This is the definition commonly used to estimate $\eta_{\oplus}$ when predicting the detection yield of future space telescopes (e.g., Kopparapu et al., 2018), and planets that meet it are referred to here and elsewhere as "exo-Earth candidates" (or EECs). However, the phrasing *potentially* habitable acknowledges that a vast array of factors affect planetary habitability, so the number of truly habitable planets per star is likely smaller than $\eta_{\oplus}$.
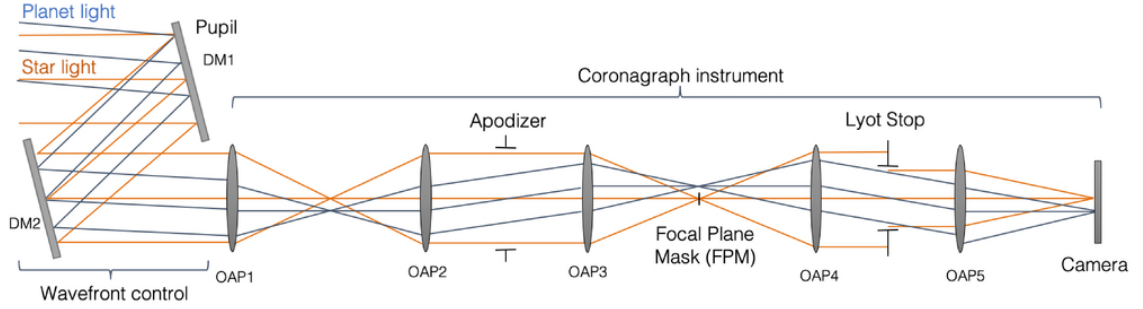
Figure 1.4: Optical layout of the coronagraphic imager for the LUVOIR space telescope concept, from Juanola-Parramon et al. (2019). Two deformable mirrors are used to correct wavefront errors, then the stellar light is removed from the image using the focal plane mask. The apodizer and Lyot stop reduce diffraction effects from the stellar light, which would otherwise overwhelm the image.

1.4 Future prospects for characterizing habitable exoplanets

Even given conservative estimates for $\eta_\oplus$, the detections by *Kepler* suggest that dozens to hundreds of Earth-sized planets exist in the habitable zones of stars within ~ 50 pc from the Sun. This finding has motivated the development of several concepts for space- and ground-based observatories and instruments that could efficiently characterize this sample and conduct the first systematic search for life beyond the solar system. Two promising observing techniques for achieving this goal are direct imaging, which measures the planet's spectrum directly by spatially resolving it from the host star, and transit spectroscopy, which measures the spectrum of the host star as filtered through the planet's atmosphere during a transit event. In either case, the spectrum contains absorption, scattering, and emission features that reveal the abundance of different atoms and molecules in the atmosphere. Here, I will review how these techniques could be implemented to study habitable planets within the coming decades, as well as the strengths, limitations, and statistical biases inherent to each. In Chapter 6, I will delve more deeply into how these factors affect the usefulness of either technique for conducting statistical studies of the terrestrial exoplanet population.
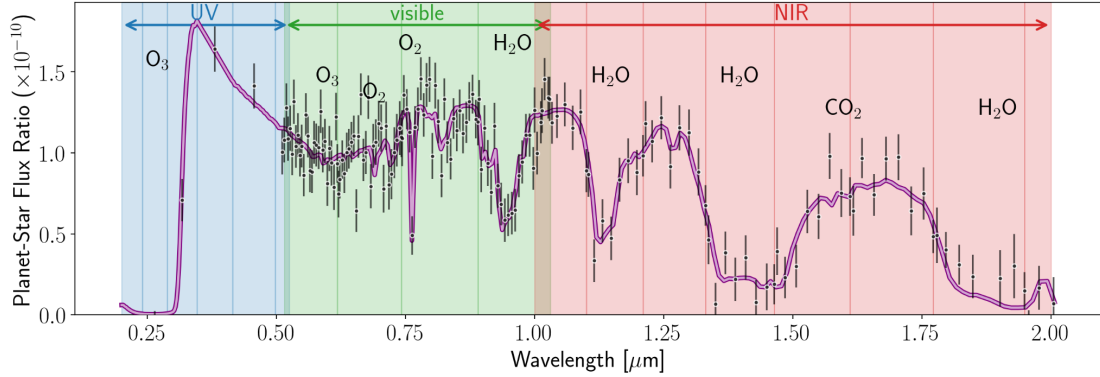
Figure 1.5: Simulated reflected light spectrum of an exo-Earth around a Sun-like star as observed with the LUVOIR space telescope, from The LUVOIR Team (2019). The imager can simultaneously observe in ultraviolet, visible, and near-infrared channels, but only one bandpass within each of these channels can be observed at a time (separated by vertical lines). Key features include H_2O (an indicator of habitability), O_2/O_3 (a potential biosignature), and CO_2.

1.4.1 Direct imaging from space

In direct imaging observations, light from the planet is spatially resolved from its host star so that the observer can measure its spectrum directly. The spectrum is dominated either by reflected light from the host star or thermal emission from the planet itself, depending on the planet's temperature and the wavelength range observed. While direct imaging is conceptually simple, it faces steep technical challenges as the star is orders of magnitude brighter than the planet. This factor is commonly quantified by the planet-to-star brightness contrast ratio, which for a planet observed in reflected light at quadrature phase (where the star-planet-observer phase angle equals 90°) can be written as:

$$\xi = \frac{A_g}{\pi} \left(\frac{R_p}{a} \right)^2 \tag{1.3}$$

where A_g is the planet's geometric albedo, R_p is its radius, and a is its separation from the host star (e.g., Traub and Oppenheimer, 2010). Since all telescopes have limited resolution and the angular planet-star separation on the sky is typically very small, the brighter light from the star overlaps with and easily obscures that of the planet. Astronomers can mitigate this effect by using large telescopes with sharper image resolution, by designing

instruments to reduce the amount of starlight reaching the detector (without also blocking light from the planet), and by developing observational and data processing techniques to efficiently subtract the stellar light from the image.

Direct imaging has been successfully employed to characterize large, intrinsically bright planets and protoplanetary and debris disks at wide angular separations from their host stars, typically corresponding to several AU (Bowler, 2016). To extend this technique to the domain of Earth-like planets will require major technological developments, and two technologies are being actively pursued to accomplish this with next-generation space telescopes (Crill and Siegler, 2017). The first is coronagraphy, in which a small occulting mask is placed within the optical path of the telescope to obscure the star's light while allowing the planet's light to pass - see Figure 1.4. This would permit observations of planets that are brighter than the coronagraph's limiting contrast ratio and that lie within a set range of angular distances from the host star, bounded by the inner and outer working angles (IWA and OWA, respectively). Both IWA and OWA are proportional to the telescope's diffraction-limited resolution (λ/D), so a larger telescope can image planets on shorter orbits or around more distant stars. Coronographic imaging of exo-Earths can only succeed if the wavefront of incident light is nearly perfectly flat, but in reality it will be distorted by imperfections in the telescope optics. A coronagraphic space telescope would therefore make use of adaptive optics - a technique originally developed to mitigate atmospheric distortion for ground-based telescopes - to correct the wavefront of light before passing it on to the instrument. To achieve the requisite near-perfect image quality for exo-Earth imaging demands exceptional thermal and pointing stability, thus driving up the spacecraft complexity and cost. Furthermore, the coronagraph can only operate over a limited wavelength range, so the full spectrum of the planet cannot be observed simultaneously and must be reconstructed from consecutive observations of narrower bandpasses (The LUVOIR Team, 2019).

The second technique to mitigate the host star's light uses a starshade, a spacecraft that flies separately to block the star light from ever reaching the telescope's mirror (Cash, 2006). The starshade likewise has an inner working angle determined by its physical extent, but there is no outer limit to its operating range. This technique conveys a few other

advantages over a coronagraph: for example, it allows the telescope to observe a much larger wavelength range simultaneously and requires comparatively looser constraints on the telescope's image quality. Those tight requirements are transferred to the starshade itself, which must launch in a compact configuration to fit in the rocket fairing, then unfurl to a diameter of tens of meters. The shape of the starshade is precisely designed to counter diffraction effects, and it must unfurl into this shape with very high accuracy. Furthermore, the necessary size of the starshade grows with the telescope's mirror diameter, making this approach less feasible with larger telescopes. Finally, the starshade must physically move from target to target, often by a distances of thousands of kilometers, and the total number of maneuvers it can perform is limited by fuel consumption. Due to this limitation, a telescope that makes use of a starshade might still possess a coronagraph with which to conduct a preliminary survey to determine which targets are worth the fuel expense and the optimal order in which to observe them (e.g., Gaudi et al., 2020).

In general, space-based direct imaging will be optimized for the study of planets in the habitable zones of nearby Sun-like stars, as the habitable zones around smaller or more distant stars typically lie within the inner working angle of the starshade or coronagraph. A directly-imaged planet's spectrum can be used to determine its atmospheric composition and preliminary surface properties (see Figure 1.5), while time-intensive multi-epoch observations can constrain its precise orbital parameters by measuring the planet's Keplerian motion. However, direct imaging cannot directly determine a planet's size or mass. As can be seen in Equation 1.3, the planet's observed brightness at a given orbital phase depends on both its size and reflectivity, which are difficult to separate through imaging observations alone. Therefore, whether a planet is an "exo-Earth candidate" as defined in Section 1.3.2 is not trivial to determine via direct imaging, but in Chapter 3, I demonstrate how a Bayesian classification scheme informed by *Kepler* statistics can greatly assist in solving this problem.

NASA has funded in-depth concept studies for two next-generation space telescopes that would directly image habitable worlds. The Large UV/Optical/Infrared Surveyor (LUVOIR; The LUVOIR Team, 2019) would have a segmented primary mirror up to 15 meters in diameter and would image planets using a coronagraphic imager with three

separate channels covering ultraviolet, optical, and near-infrared bandpasses. The smaller Habitable Exoplanet Observatory (HabEx; Gaudi et al., 2020) would have a monolithic 4-meter diameter mirror and would image planets using a separate starshade in addition to a multiple-channel coronagraphic imager. Of the two, LUVOIR would observe an overall larger sample of planets, due to its better resolution and light-collecting area, but the requisite size and complexity of the spacecraft is likely to entail a considerable cost. The HabEx concept aims for a simpler spacecraft, but the cost and complexity of the starshade itself could overshadow that advantage. The number of exo-Earth candidates detectable by a coronagraphic space telescope scales with mirror diameter approximately as $D^{1.8}$ (Stark et al., 2014), as increasing the diameter decreases the IWA, allowing the telescope to probe habitable zones around more distant stars. However, the construction of even a 15-meter space telescope with the image quality required for direct imaging remains a formidable task, so the predicted detection yield of the LUVOIR telescope ($\sim$ 700 planets including $\sim$ 54 exo-Earth candidates, assuming $\eta_\oplus = 24\%$) likely represents the maximum number of planets that can be observed through direct imaging within the next few decades.

1.4.2 Transit spectroscopy from space

Transit spectroscopy is an extension of the transit technique used by *Kepler* to discover and measure the sizes of transiting exoplanets. In it, the star is observed with a spectrograph during the transit event such that the transit depth can be measured as a function of wavelength (known as the transit spectrum; see Figure 1.6). As a small portion of the host star's light is filtered through the planet's atmosphere, the transit spectrum traces the wavelength-dependent optical depth of the atmosphere due to scattering and absorption effects, and from it one can retrieve the abundance of atmospheric species that produce them. To first order, the fraction of stellar light blocked by the planet's atmosphere above the continuum (or "white-light") transit depth due to the planet's radius is equal to the fraction of the stellar disk occulted by one atmospheric scale height:

$$\Delta\delta \sim (R_p/R_*)^2(h/R_p) \qquad (1.4)$$
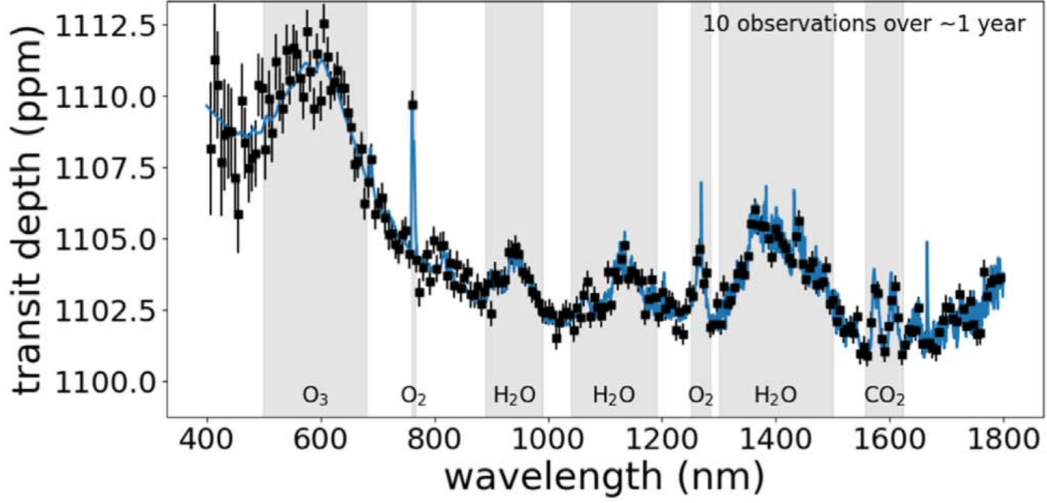
Figure 1.6: Simulated transit spectrum for an exo-Earth orbiting a nearby mid-M dwarf as observed with the *Nautilus* Space Observatory, from Apai et al. (2019). This spectrum represents the combined result of ten separate transit observations over the course of a year - however, it does not account for the impact of cloud cover which could reduce the strength of absorption features by a factor of several.

where h is the atmospheric scale height and R_p and R_* are the planet and stellar radii (e.g., Winn, 2010). The actual value of $\Delta\delta$ as a function of wavelength depends on the wavelength-dependent opacity of the atmosphere, and high-altitude clouds can serve to greatly dilute $\Delta\delta$ as they set a higher white-light depth.

To date, transit spectroscopy has been the most widely applied technique for studying exoplanet atmospheres - mostly those of short-period gas giants whose large sizes, frequent transits, and hot inflated atmospheres make them ideal candidates for this technique. The success of transit spectroscopy can be attributed in part to its low technological threshold, as it requires only a moderate resolution spectrograph and a telescope with sufficient light-collecting area. As such, the Hubble Space Telescope has been used to characterize dozens of planetary atmospheres despite not being designed with this goal in mind (e.g., Sing et al., 2016). However, while Hubble has been used to conduct preliminary atmospheric characterizations of planets in the habitable zone (de Wit et al., 2018; Benneke et al., 2019), a thorough search for biomarkers remains out of its reach due primarily to its limited collecting area. Even telescopes capable of transit spectroscopy cur-

rently planned to launch in the future, such as the James Webb Space Telescope (JWST), will at best succeed at characterizing only a small number of habitable worlds (Lustig-Yaeger et al., 2019b). In order to search for life on transiting planets, dedicated space telescopes with large light-collecting areas must be devised.

Since its measurement precision is limited by its target's brightness, transit spectroscopy scales especially well with the observatory's light-collecting area. To first order, a four-fold increase in area enables a two-fold increase in the distance from the Sun to which habitable planets can be characterized, and therefore an eight-fold increase in the achievable sample size (which is proportional to the volume probed). Furthermore, observations across multiple telescopes or consecutive transit events can be digitally combined to effectively increase the light-collecting power. Transit spectroscopy is optimized for planets that orbit stars less massive than the Sun, in part because the atmospheric signal (Equation 1.4) is $\sim 100\times$ larger due to the smaller stellar radius, and in part because planets in their habitable zones have shorter orbital periods, transiting multiple times per month as opposed to once per year. This bias toward low-mass stars is favorable since, as discussed in Section 3.3.2, $\eta_\oplus$ is likely much higher for these systems.

However, transit spectroscopy faces two major astrophysical obstacles that each could jeopardize its usefulness for characterizing habitable worlds. First, it is difficult to distinguish spectral signatures due to the planet's atmosphere from those produced by active regions (cool sunspots and hot faculae) on the star itself. This effect has already been observed in transit spectra of Earth-sized planets (Zhang et al., 2018), and could overwhelm the signal produced by biomarkers in the atmospheres of habitable worlds orbiting low-mass stars (Rackham et al., 2018). Second, high-altitude clouds and hazes in a planet's atmosphere can obscure molecular and atomic absorption, an effect also observed in transit spectra of gas giants (e.g., Kreidberg et al., 2014). While planets around low-mass stars are better targets for transit spectroscopy, such planets are likely to be tidally-locked to their host stars, leading to permanent day and night sides (Barnes et al., 2016). General circulation models have predicted clouds to exist in greater abundance and at higher altitudes above these planets compared to Earth due to more efficient convection in the dayside atmosphere (e.g., Komacek and Abbot, 2019). This would greatly reduce the

strength of absorption features produced by species below the clouds (e.g., Fauchez et al., 2019). In Chapter 6, I find that the impact of clouds predicted by such models could lead to an order of magnitude decline in the number of potentially habitable worlds accessible through transit spectroscopy, although this impact may be lessened if the cloud covering fraction is lesser for some planets or even during subsequent observations of individual planets.

If these obstacles can be overcome, its favorable scaling with light-collecting area means that space-based transit spectroscopy is likely the only technique capable of characterizing the atmospheres of hundreds of exo-Earth candidates within the next few decades. As with direct imaging, multiple concepts have been developed for space telescopes that could characterize habitable planets in this way. LUVOIR would possess the light-collecting area to study potentially ~ 10 transiting Earth-like planets around the nearest low-mass stars, thus complementing its direct imaging survey of Sun-like stars (The LUVOIR Team, 2019). To exceed this sample size will require a much greater light-collecting area, but the cost of space telescopes scales formidably with mirror diameter (Stahl, 2019). The *Nautilus* concept (Apai et al., 2019) aims to circumvent this cost barrier by leveraging a large ($\sim$8.5 m diameter), lightweight multi-order diffractive element, or MODE lens (Milster et al., 2020). This lens would enable a lighter telescope with a simpler optical path and an order of magnitude greater tolerance to optical misalignment than a comparably-sized mirror. Furthermore, the telescope would be designed to be replicable at an affordable cost, enabling the production of dozens of identical units whose combined light-collecting area would match that of a single 50-meter diameter telescope. In an optimistic scenario, such an array could probe the atmospheres of as many as one thousand exo-Earth candidates (Apai et al., 2019).

1.4.3 Ground-based observatories

The coming decade will see the commissioning of optical/infrared telescopes with primary mirrors wider than twenty meters in diameter, collectively referred to as extremely large telescopes (ELTs). ELTs currently planned or under construction include the Giant Magellan Telescope (GMT; Johns et al., 2012), European Extremely Large Telescope

(E-ELT; Gilmozzi and Spyromilio, 2007), and Thirty Meter Telescope (TMT; Sanders, 2014). In the past, the image resolution of ground-based telescopes has been fundamentally limited by atmospheric distortion. This problem has been greatly reduced by advances in adaptive optics technology over the past decade (Close, 2016), meaning that upcoming ELTs equipped with high-contrast imagers will have even finer image resolution than planned direct imaging telescopes such as LUVOIR, allowing them to observe planets at narrower separations from their stars. On the other hand, they are unlikely to achieve the same level of thermal and mechanical stability as a space telescope, leading to a higher planet-star contrast limit. This pair of circumstances makes them ideal for observing temperate terrestrial planets orbiting nearby low-mass stars, where the habitable zone is closer in and thus ξ is much larger (Equation 1.3). First steps have already been taken in this direction, as the Very Large Telescope (VLT) in Chile was recently used to search for temperate worlds in the Alpha Centauri system, demonstrating sensitivity to sub-Neptune-sized planets within the habitable zone (Wagner et al., 2021).

With their formidable light-collecting areas, ELTs could prove useful for transit spectroscopy of habitable planets as well. Ground-based telescopes are now routinely used to produce high-quality transit spectra of gas giants; my work in Chapter 4 demonstrates one such study. However, ground-based transit spectroscopy of the atmospheres of habitable worlds will be inhibited by the fact that the molecules of greatest interest - for example, H_2O and O_2 - also produce overlapping absorption features in Earth's atmosphere. Fortunately, recent studies have shown that high-resolution spectroscopy can potentially counter this problem. Since a transiting planet's orbital parameters are measurable, the Doppler shift of its absorption lines can be calculated, and this information can be used to distinguish those lines from Earth's in a high-resolution spectrum. This technique could be applied to detect features such as O_2 in the atmospheres of transiting exo-Earth candidates using ground-based ELTs (Snellen et al., 2013).

1.5   Thesis motivation and outline

The discoveries of *Kepler* have set the stage for the first systematic search for habitable planets and life beyond the solar system, and it is likely that at least one space-based observatory that can conduct this search will begin development within the near future. This exciting possibility raises, in turn, important questions about how to proceed with the effort. First: how can we translate limited data about a planet's basic properties and rudimentary atmospheric composition into meaningful constraints on the planet's habitability and the possible presence of life? Second: given limited resources, which observing techniques and which specific observatory concepts are most likely to yield the maximum science return? And finally: how can we combine observations of individual planets to form a more complete picture of what makes planets habitable and what factors give rise to life?

In this thesis, I propose strategies for answering these questions based on exoplanet statistics and probabilistic reasoning. First, I demonstrate how Bayesian techniques based on *Kepler* statistics can be used to more confidently constrain the habitability of planets for which limited information is available. In Chapter 2, I apply this approach to determine the likely composition of Proxima Centauri b, while in Chapter 3, I show how a similar technique could be used to optimize the survey strategy of space telescopes designed to directly image habitable planets. In Chapter 4, I present my analysis of the transit spectrum of the hot Jupiter WASP-4b, in which I use a Bayesian approach to quantify evidence for stellar contamination in the spectrum. Similar techniques may be necessary to properly characterize habitable exoplanets through transit spectroscopy in the future. In Chapter 5, I show that by leveraging a statistical approach to exoplanet characterization, future space observatories could study how habitable planets evolve over time in comparison with Earth. In Chapter 6, I expand upon this concept by developing a general framework to evaluate the ability of future space observatories to test similar hypotheses about the formation and evolution of habitable worlds. Finally, in Chapter 7, I summarize my key findings and their implications for the development of new technologies to search for life beyond the solar system.

CHAPTER 2

Probabilistic Constraints on the Mass and Composition of Proxima b

This chapter has been published as Bixel and Apai (2017) and is reproduced here with permission.

Abstract

Recent studies regarding the habitability, observability, and possible orbital evolution of the indirectly detected exoplanet Proxima b have mostly assumed a planet with $M \sim 1.3$ $M_\oplus$, a rocky composition, and an Earth-like atmosphere or none at all. In order to assess these assumptions, we use previous studies of the radii, masses, and compositions of super-Earth exoplanets to probabilistically constrain the mass and radius of Proxima b, assuming an isotropic inclination probability distribution. We find it is $\sim 90\%$ likely that the planet's density is consistent with a rocky composition; conversely, it is at least 10% likely that the planet has a significant amount of ice or an H/He envelope. If the planet does have a rocky composition, then we find expectation values and 95% confidence intervals of $\langle M \rangle_{\mathrm{rocky}} = 1.63^{+1.66}_{-0.72}\ M_\oplus$ for its mass and $\langle R \rangle_{\mathrm{rocky}} = 1.07^{+0.38}_{-0.31}\ R_\oplus$ for its radius.

2.1 Introduction

The recent radial velocity detection of a planet in the habitable zone of the nearby M dwarf Proxima Centauri (hereafter 'Proxima b' and 'Proxima') (Anglada-Escudé et al., 2016) has spurred over a dozen theoretical papers speculating on the planet's atmosphere (e.g., Brugger et al., 2016; Goldblatt, 2016), habitability (e.g., Ribas et al., 2016; Turbet et al., 2016), and orbital and formation histories (e.g., Barnes et al., 2016; Coleman et al., 2017) as well as prospects for a direct detection or atmospheric characterization (e.g., Lovis et al., 2017; Luger et al., 2017). As Proxima is the nearest neighbor to the solar system, it has been suggested as a target for future space missions, including those hoping to characterize its atmosphere and search for life (e.g., Belikov et al., 2015; Schwieterman et al., 2016).

In many of these studies, authors have assumed a rocky planet with a thin atmosphere or no atmosphere at all, and some have assumed a mass near or equal to the projected mass of $M \sin(i) = 1.27^{+0.20}_{-0.17} \, M_\oplus$, but little has been done to assign a degree of certainty to these assumptions. Most notably, previous studies have revealed two distinct populations of exoplanets with super-Earth radii: 'rocky' planets composed almost entirely of rock, iron, and silicates with at most a thin atmosphere, and 'sub-Neptune' planets which must contain a significant amount of ice or a H/He envelope (e.g., Rogers, 2015; Weiss and Marcy, 2014). If there is a significant probability that Proxima b is of the latter composition, then this should be taken into account when assessing its potential habitability or observability.

In this letter, we generate posterior distributions for the mass of Proxima b using Monte Carlo simulations of exoplanets with an isotropic distribution of inclinations, where the radii, masses, and compositions of the simulated planets are constrained by results from combined transit and radial velocity measurements of previously detected exoplanets. By comparing the posterior mass distribution to the composition of planets as a function of mass, we determine the likelihood that Proxima b is, in fact, a rocky world with a thin (if any) atmosphere.

2.2 Prior assumptions

Radial velocity and transit studies of exoplanets have yielded mass and radius measurements for a statistically significant number of targets, thereby enabling the study of how the occurrence and composition of exoplanets varies with planet radii, orbital periods, and host star type. In this section, we review previous results, which we will use to place stronger constraints on the mass and composition of Proxima b.

2.2.1 $\sin(i)$ distribution

It can be shown (e.g., Ho and Turner, 2011) that the probability distribution of $\sin(i)$ corresponding to an isotropic inclination distribution is

$$P(\sin(i)) = \sin(i)/\sqrt{1-\sin^2(i)} \tag{2.1}$$

Since this distribution peaks at $\sin(i) = 1$, the mass distribution of an RV-detected planet - assuming no prior constraints on the mass - peaks at the minimum mass M_0.

In their models of the possible orbital histories of Proxima b, Barnes et al. (2016) find that galactic tides could have inflated the eccentricity of the host star's (at the time unconfirmed) orbit around the α Cen binary, leading to encounters within a few hundred AU and the possible disruption of Proxima's planetary system. If so, this could affect the likely inclination of the planet in a non-isotropic way. However, Kervella et al. (2017) have presented radial velocity measurements showing that Proxima is gravitationally bound to the α Cen system with an orbital period of 550,000 years, an eccentricity of ~ 0.5, and a periapsis distance of 4,200 AU. At this distance, the ratio of Proxima's gravitational field to that of α Cen at the planet's orbit (~ 0.05 AU) is greater than 10^8; unless Proxima's orbit was significantly more eccentric in the past, it seems unlikely that α Cen would have disrupted the system.

2.2.2 Occurrence rates for M dwarfs

Mulders et al. (2015b) provide up-to-date occurrence rates of planets around M dwarf stars from the *Kepler* mission. The sample is limited to $2 < P < 50$ days, over which they

find the occurrence rates to be mostly independent of the period. The binned rates and a regression curve, as well as their uncertainties, are presented in Figure 2.1.

Kepler statistics for M dwarfs remain incomplete below 1 $R_\oplus$, but complete statistics for earlier-type stars suggest a flat distribution for $0.7 < R < 1$ $R_\oplus$ (Mulders et al., 2015b). Since mass-radius relationships typically find a strong dependence of mass on radius ($M \propto R^{3-4}$) (e.g. Weiss and Marcy, 2014; Wolfgang et al., 2016), we assume *a priori* that Proxima b ($M \gtrsim 1.3$ $M_\oplus$) is larger than 0.7 $R_\oplus$. Therefore, for this letter we adopt the regression curve fitted to the binned data, but set the occurrence rates to be flat for $R < 1$ $R_\oplus$.

2.2.3 Compositions

Multiple works (e.g. Weiss and Marcy, 2014; Rogers, 2015) have determined the existence of two distinct populations of exoplanets smaller than Neptune ($R \lesssim 4$ $R_\oplus$): a small radius population with densities consistent with an entirely iron and silicate composition (hereafter 'rocky'), and a large radius population with lower density planets that must have significant amounts of ice or a thick H/He atmosphere (hereafter 'sub-Neptunes').

Rogers (2015) studies the abundance of planets of each composition as a function of radius. They define $f_\alpha(R)$ as the likelihood that a planet of radius R is dense enough to be consistent with a rocky composition, and determine f_α for a sample of planets with known masses and radii. They suggest fitting the data with a two-parameter linear model:

$$f_\alpha(R_P, R_{\text{thresh}}, \Delta R)$$

$$= \begin{cases} 1 & R_P < R_{\text{thresh}} - \frac{1}{2}\Delta R \\ 0.5 + \dfrac{R_{\text{thresh}} - R_P}{\Delta R} & |R_P - R_{\text{thresh}}| < \frac{1}{2}\Delta R \quad (2.2) \\ 0 & R_P > R_{\text{thresh}} + \frac{1}{2}\Delta R \end{cases}$$

They find a step function to best describe the data, with ΔR fixed at zero and $R_{\text{thresh}} \approx$ 1.5 $R_\oplus$. For the purposes of this letter, we prefer this fit, but will also vary R_{thresh} and ΔR to see how they affect our results.
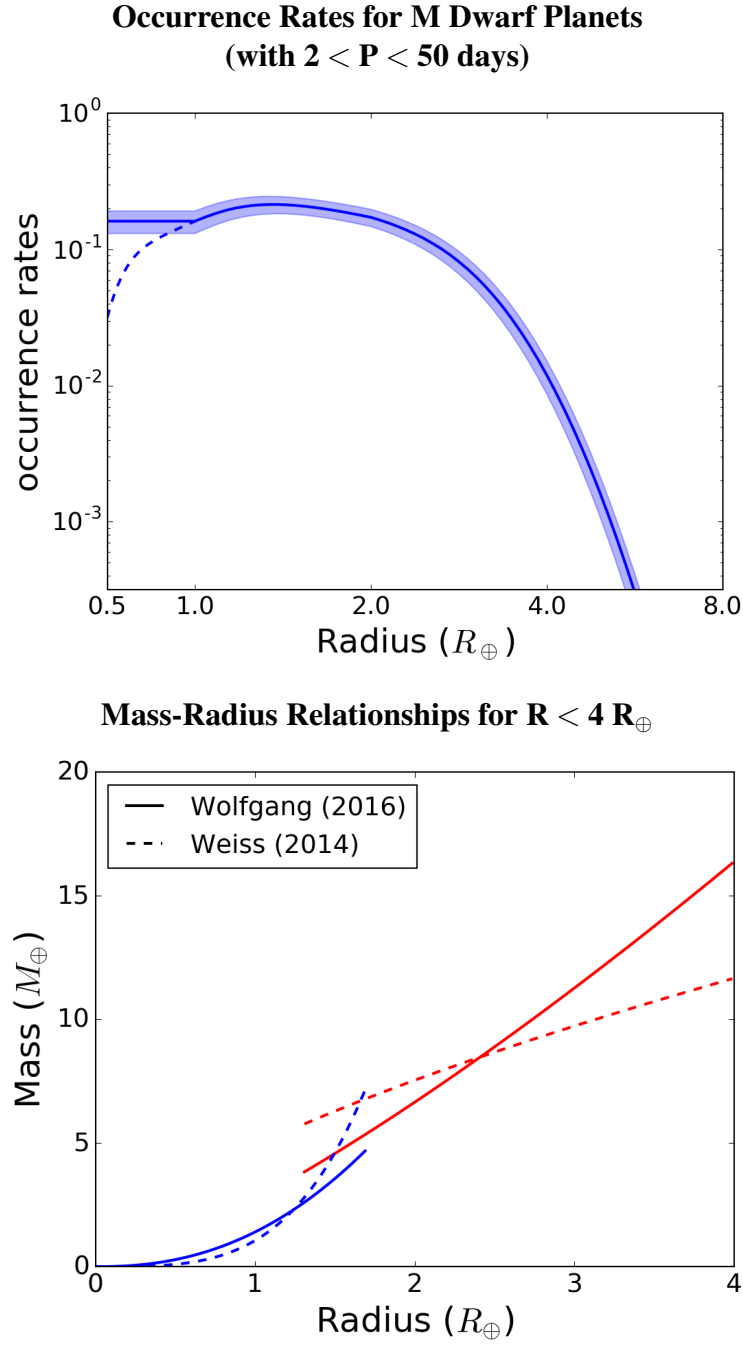
Figure 2.1: Top: Occurrence rates from Mulders et al. (2015b), fitted by a regression curve. We adjust the rates below 1 $R_\oplus$ (dotted) to be flat, since the sample is incomplete in this range. Bottom: Mass-radius relationships for the rocky (blue) and sub-Neptune (red) populations. The plotted relationships are from Wolfgang et al. (2016) (solid) and Weiss and Marcy (2014) (dashed).

We stress that a planet for which $f_\alpha = 1$ is only *sufficiently* dense to be rocky; we still cannot necessarily exclude an ice or volatile component. Here, we will assume that all planets for which $f_\alpha = 1$ follow the low-radius M-R relationships given in the following section, which were empirically fitted without prior knowledge of the planets' compositions. For simplicity, we refer to these as 'rocky' planets, and the other population as 'sub-Neptunes', but we will revisit this distinction later on.

Since Proxima b is in the habitable zone, it receives an amount of stellar flux comparable to that received by Earth, so we should bear in mind the possibility that the volatile envelope of a sub-Neptune could be lost due to photoevaporation. Owen and Mohanty (2016) model rocky planets with thick H/He envelopes in the habitable zones of M dwarfs, finding that planets with $M > 0.8\ M_\oplus$ maintain their envelopes over Gyr timescales and are therefore uninhabitable. The 2σ lower limit on the minimum mass of Proxima b is $0.93\ M_\oplus$, so it is unlikely that any H/He envelope on the planet would evaporate under this rule. However, we note that this study focuses on planets with a primarily rocky composition, so it may not be directly applicable to habitable zone sub-Neptunes.

Additionally, Zahnle and Catling (2013) empirically define boundaries for atmospheric evaporation as a function of stellar heating, escape velocity, and atmospheric composition. In particular, a planet receiving an Earth-like flux must have an escape velocity above ~ 8 km/s in order to maintain an H_2 atmosphere for 5 Gyr. We will revisit this requirement in Section 2.4.2.

2.2.4   Mass-radius relationships

Empirically derived relationships between exoplanet masses and radii rely on radial velocity (RV) or transit-timing variation (TTV) measurements of transiting exoplanet masses. Weiss and Marcy (2014) fit a mass-radius (hereafter M-R) relationship to a sample of 65 transiting exoplanets, in which they find evidence for the two populations discussed in Section 2.2.3. Through least-squares regression, they find the densities of the rocky planets to increase linearly with planet radius:

$$\rho_P = 2.43 + 3.39 \left(\frac{R_P}{R_E} \right) \text{g cm}^{-3} \tag{2.3}$$

while the RV-measured masses of sub-Neptunes increase nearly linearly with planet radius:

$$\frac{M_P}{M_\oplus} = 4.87 \left(\frac{R_P}{R_E}\right)^{0.63} \tag{2.4}$$

Wolfgang et al. (2016) use an expanded version of this data set to fit power law M-R relationships using a more statistically robust Bayesian method. For the rocky planets, they find

$$\frac{M_P}{M_\oplus} = 1.4 \left(\frac{R_P}{R_E}\right)^{2.3} \tag{2.5}$$

and for the sub-Neptunes with RV-measured masses,

$$\frac{M_P}{M_\oplus} = 2.7 \left(\frac{R_P}{R_E}\right)^{1.3} \tag{2.6}$$

Due to the larger sample size and more robust fitting procedure, we adopt Equations 2.5 and 2.6 as our preferred M-R relationships, but for completeness we consider the Weiss and Marcy (2014) relationships as well. We find that the choice of M-R relationships has a minimal impact on our final results. Both sets of relationships are plotted in Figure 2.1.

It is important to note that the above relationships for sub-Neptunes exclude masses measured by TTV, since TTV masses have been found to be systematically lower than RV masses. This could indicate a selection bias or systematic error in the method used, but since Proxima b was detected through RV measurements, we believe it is proper to exclude the TTV masses either way.

It is also clear that there is a significant spread in the masses of the observed planets. Wolfgang et al. (2016) suggest a spread of $\pm 1.9\ M_\oplus$ for the sub-Neptune planets, which we adopt for our simulations. For rocky planets, the spread is noticeably smaller. There are too few planets to constrain this spread, but it should most likely increase with mass, so we arbitrarily define the spread to be 30% of the calculated mass.

2.3 Method

2.3.1 Simulated sample

The fitted occurrence rates and their uncertainties ($f \pm df$) are given in even bins in log-space. We use them to generate a random sample of radii, where the number of radii in each bin (r_0) is selected from a normal distribution with mean value $f(r_0)$ and standard deviation $df(r_0)$. We find that the results converge for 1,000 samples of the occurrence rates, with each sample containing $\sim 10^6$ radii.

To each radius, we assign a composition ('rocky' or 'sub-Neptune') based on the model of Rogers (2015) (Equation 2.2), with $R_{\text{thresh}} = 1.5\,R_\oplus$ and $\Delta R = 0$. We then assign a mass to each radius and composition from a Gaussian distribution with mean value $M(R)$ - calculated using our chosen M-R relationships (Equations 2.5 and 2.6) - and a standard deviation dM that represents the spread. We choose a spread proportional to the calculated mass for rocky planets ($dM = 0.3 \times M(R)$), but a constant spread for sub-Neptunes ($dM = 1.9\,M_\oplus$). We also reject negative masses, which could in principle bias the assigned masses towards higher-than-average values - however, we find that only a negligible number of masses are rejected.

Finally, we assign a line-of-sight inclination parameter $\sin(i)$ to each planet, drawn from the isotropic probability distribution discussed in Section 2.2.1.

2.3.2 Prior and posterior probability distributions

The prior mass and radius distributions, $P(M)$ and $P(R)$, can be derived directly from the simulated sample. Factoring in the projected minimum mass M_0, the posterior distributions $P(M|M_0)$ and $P(R|M_0)$ can be calculated from Bayes' formula:

$$P(X|M_0) = \frac{P(M_0|X)P(X)}{P(M_0)} \tag{2.7}$$

where X represents mass or radius. Since M_0 is known, $P(M_0)$ is just a normalizing constant. Taking $M_0 = 1.27^{+0.20}_{-0.17}\,M_\oplus$ as the projected mass of Proxima b and the upper limit $\sigma_{M_0} = 0.20\,M_\oplus$ as its standard deviation, we calculate for each simulated planet

$$P_i(M_0|X) = \exp\left(-(M_0 - M_i \sin_i(i))^2/2\sigma_{M_0}^2\right) \tag{2.8}$$

Then $P(M_0|M)$ and $P(M_0|R)$ are the average values of $P_i(M_0|X)$ for each bin in mass or radius. The prior and posterior distributions are calculated for each sample of 10^6 planets, and the final results are taken to be the mean result of 1,000 samples.

2.3.3   Posterior compositional probability

The prior probability that a planet in a given mass bin is rocky is equal to the number of simulated rocky planets in that bin divided by the total number of planets in the same bin. Since we want to know the likelihood that Proxima b belongs to the 'rocky' population, we multiply this prior composition probability distribution by the posterior mass distribution from the previous section and integrate over all masses.

2.4   Results

2.4.1   Mass distributions

The prior and posterior mass probability distributions for Proxima b are plotted in Figure 2.2. The shaded regions demonstrate the relative contributions of the populations at each mass. The prior distribution is valid for RV-detected planets around M dwarfs with intermediate periods ($2 < P < 50$ days) and radii ($0.7 < R < 4\ R_\oplus$), while the posterior distribution can be taken as the mass probability distribution for Proxima b.

For reference, we include the posterior distribution given *no* prior constraints on the mass; that is, the distribution resulting from an isotropic $\sin(i)$ distribution and the measured M_0 with its uncertainty. We find that this nearly matches our result, since both $P(\sin(i))$ and $P(M)$ are bottom-heavy.

Figure 2.3 shows the cumulative probability that $M < X$ for both of the considered M-R relationships (Section 2.2.4) as well as for the case of no prior mass distribution. We find that there is little difference between the results for each M-R relationship.

2.4.2   Escape velocity

In order to verify that sub-Neptune planets can maintain H_2 envelopes in the habitable zone, we compare the escape velocities of our simulated sub-Neptunes to the ~ 8 km/s

Prior mass distribution

Posterior mass distribution
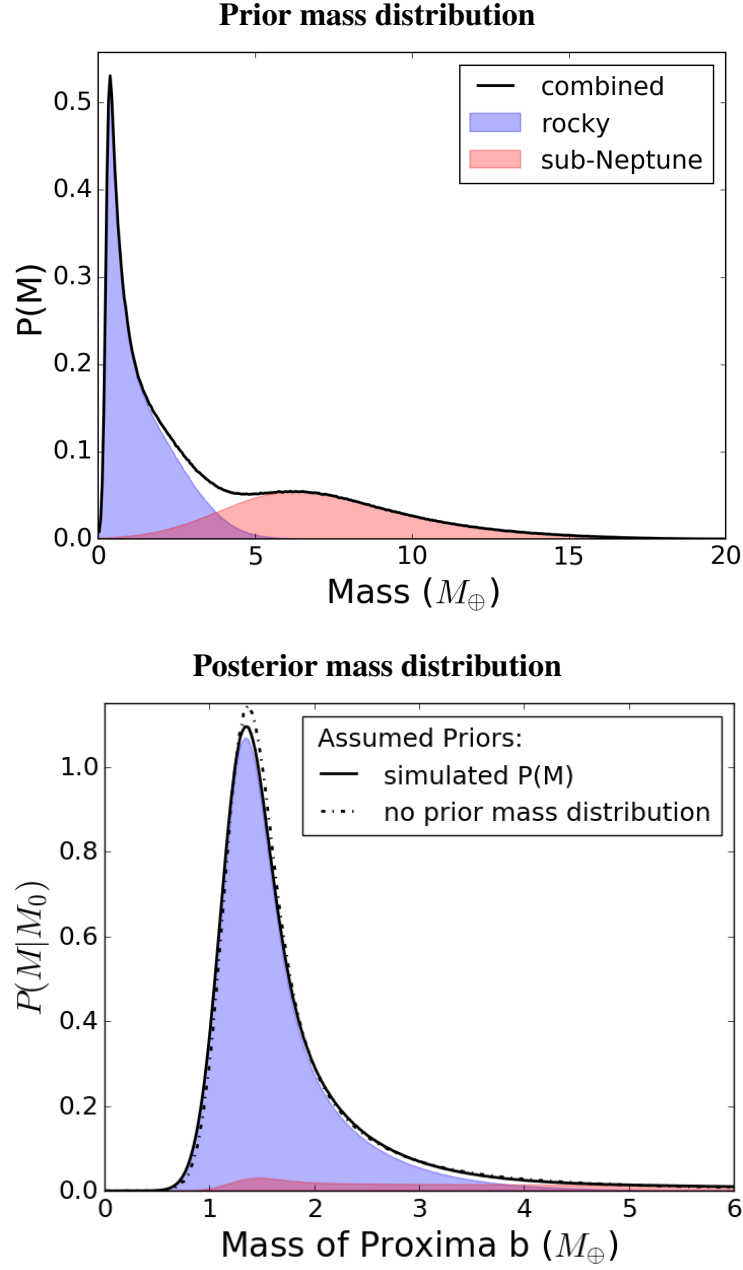
Figure 2.2: Prior (top) and posterior (bottom) mass distributions for the simulated sample. The blue and red shaded regions represent contributions due to rocky and sub-Neptune planets, respectively. The dash-dotted line is the posterior distribution assuming a flat prior distribution. The binning is 0.01 $M_\oplus$.

Cumulative mass probability distribution
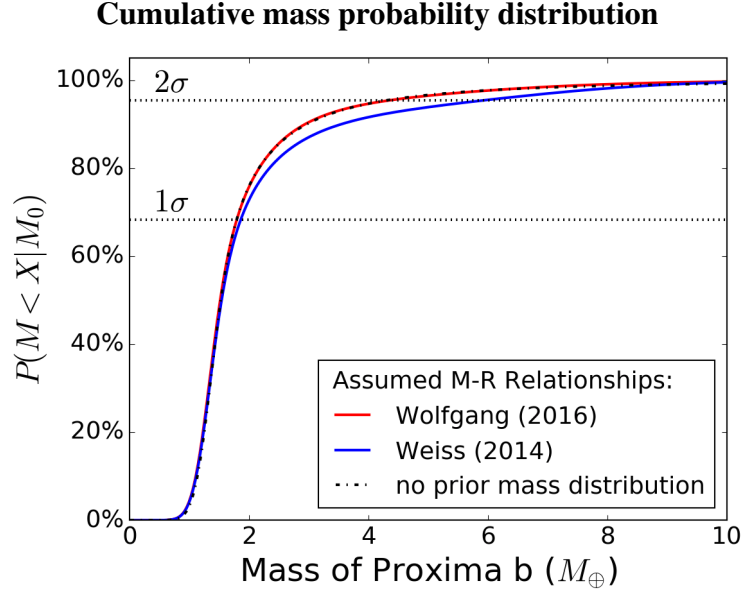
Figure 2.3: The cumulative mass probability distribution for our simulated posterior mass distribution (solid) and assuming a flat prior $P(M)$ (dashdot). The dotted lines intersect 68% and 95% confidence upper limits on the mass.

cutoff for H_2 atmospheric escape (assuming an Earth-like stellar flux) defined by Zahnle and Catling (2013). In both the prior and posterior distributions of escape velocities, we find that fewer than 1% of the sub-Neptunes have escape velocities below this threshold, with most having $v_e \gtrsim 15$ km/s. Therefore, we do not believe that Proxima b will be subject to significant atmospheric loss if it has a sub-Neptune composition.

2.4.3 Composition

Table 2.1 lists the sets of parameters for which we run the simulation, including the mass spread dM for each composition and the central value (R_{thresh}) and width (ΔR, if nonzero) of the transition region defined by Equation 2.2. The following results for each case are given: the probability P_{rocky} that Proxima b belongs to the 'rocky' category of planets, i.e. that its density is *consistent* with a fully iron and silicate composition, and the expectation values $\langle M \rangle_{\text{rocky}}$ and $\langle R \rangle_{\text{rocky}}$ of the mass and radius under the assumption that it belongs to this population.

Case A is most consistent with the previous work we have cited, so we take it as our

primary result. In this case, there is a $\sim 90\%$ probability that Proxima b belongs to the 'rocky' population, with an $\sim 10\%$ likelihood that it belongs to the 'sub-Neptune' population. In the case that it is rocky, the expectation values (and 95% confidence intervals) for the mass and radius are $\langle M \rangle_{\rm rocky} = 1.63^{+1.66}_{-0.72}\ M_\oplus$ and $\langle R \rangle_{\rm rocky} = 1.07^{+0.38}_{-0.31}\ R_\oplus$.

We investigate the effect of increasing (Case B) and decreasing (Case C) the mass spread for each composition, which results in lower and higher values of $P_{\rm rocky}$, respectively. This results from low-radius ($R \sim 1.5\ R_\oplus$), low-mass sub-Neptunes; when dM is large, they can lie significantly below the M-R relation with masses between 1 and 2 $M_\oplus$, so that they are indistinguishable from the rocky planets in the mass domain.

In Cases D and E, we determine the effect of raising or lowering the threshold radius $R_{\rm thresh}$ at which the rocky and sub-Neptune populations are split. A 0.2 $R_\oplus$ offset in either direction, which encompasses most of the values suggested in the literature, results in a $\sim 5\%$ to 8% shift in $P_{\rm rocky}$, where higher threshold radii allow for more rocky planets and therefore a higher probability of a rocky composition. Furthermore, allowing for a non-zero width ΔR to the cutoff region allows sub-Neptunes to exist with lower radii and masses, thereby decreasing $P_{\rm rocky}$.

In all cases, we find $P_{\rm rocky}$ to be between 80% and 95% using the Wolfgang et al. (2016) M-R relationship, and we find similar values using the Weiss and Marcy (2014) relationship (e.g. $P_{\rm rocky} = 90.7\%$ for Case A), so this result does not vary substantially over the range of reasonable values for the input parameters.

2.5   Conclusions

By considering occurrence rates from the *Kepler* mission and empirically derived M-R relationships, we derive a posterior probability distribution for the actual mass of Proxima b. If the planet has a rocky composition, i.e. if it obeys the low-radius M-R relationship of Wolfgang et al. (2016), then the expectation values of the mass and radius (with 95% confidence intervals) are $\langle M \rangle_{\rm rocky} = 1.63^{+1.66}_{-0.72}\ M_\oplus$ and $\langle R \rangle_{\rm rocky} = 1.07^{+0.38}_{-0.31}\ R_\oplus$.

In all of our simulations, we find a probability of 80% to 95% that Proxima b belongs to the 'rocky' population of planets defined in Section 2.2.3. In our 'best guess' scenario

(Case A), this probability is 90%. Critically, we note that we have assumed all planets with $f_\alpha = 1$ (according to the Rogers (2015) criterion) are rocky planets, while in reality their density is only consistent with such a composition. With this in mind, it is safest to say that there is *at least* a 10% chance that Proxima b has a sub-Neptune composition. If it is a sub-Neptune, then its surface gravity is high enough that it could maintain a thick hydrogen atmosphere.

For future theoretical work involving the habitability and detectability of Proxima b, we advise caution regarding assumptions made about its mass or composition; if Proxima b does possess a thick H/He envelope, then it is likely not habitable in the traditional sense. Even if the mass could be further constrained, sub-Neptunes have been measured with masses as low as $\sim 1\ M_\oplus$, so the composition cannot be conclusively inferred from the mass alone. Nevertheless, the rocky composition originally asserted by Anglada-Escudé et al. (2016) remains the most likely possibility.

2.5.1 Acknowledgements

Table 2.1.   Monte-Carlo Simulation Parameters and Results

Case	Parameters				Results		
	dM (rocky)	dM (sub-Neptune)	R_{thresh}	ΔR	P_{rocky}	$\langle M \rangle_{\text{rocky}}$	$\langle R \rangle_{\text{rocky}}$
Case A	$\mathbf{0.3 \times M}$	$\mathbf{1.9\ M_\oplus}$	$\mathbf{1.5\ R_\oplus}$	**-**	**89.9%**	$\mathbf{1.63^{+1.66}_{-0.72}\ M_\oplus}$	$\mathbf{1.07^{+0.38}_{-0.31}\ R_\oplus}$
Case B	$0.6 \times M$	$3.8\ M_\oplus$	$1.5\ R_\oplus$	-	84.6%	$1.65^{+1.95}_{-0.73}\ M_\oplus$	$1.03^{+0.42}_{-0.36}\ R_\oplus$
Case C	$0.1 \times M$	$0.7\ M_\oplus$	$1.5\ R_\oplus$	-	93.6%	$1.65^{+1.52}_{-0.73}\ M_\oplus$	$1.06^{+0.36}_{-0.24}\ R_\oplus$
Case D	$0.3 \times M$	$1.9\ M_\oplus$	$1.7\ R_\oplus$	-	94.6%	$1.71^{+2.13}_{-0.79}\ M_\oplus$	$1.10^{+0.50}_{-0.33}\ R_\oplus$
Case E	$0.3 \times M$	$1.9\ M_\oplus$	$1.3\ R_\oplus$	-	81.6%	$1.52^{+1.15}_{-0.62}\ M_\oplus$	$1.02^{+0.26}_{-0.27}\ R_\oplus$
Case F	$0.3 \times M$	$1.9\ M_\oplus$	$1.5\ R_\oplus$	$1.2\ R_\oplus$	84.8%	$1.64^{+1.99}_{-0.73}\ M_\oplus$	$1.06^{+0.53}_{-0.30}\ R_\oplus$
Case G	$0.6 \times M$	$3.8\ M_\oplus$	$1.5\ R_\oplus$	$1.2\ R_\oplus$	81.1%	$1.65^{+2.13}_{-0.75}\ M_\oplus$	$1.02^{+0.63}_{-0.36}\ R_\oplus$

Note. — The resulting values of P_{rocky}, $\langle M \rangle_{\text{rocky}}$, and $\langle R \rangle_{\text{rocky}}$ for different mass spreads dM and compositional parameters R_{thresh} and ΔR. The expectation values are reported with 95% confidence intervals.

CHAPTER 3

Identifying Exo-Earth Candidates in Direct Imaging Data through Bayesian
Classification

*This chapter has been published as Bixel and Apai (2020a) and is reproduced here with
permission.*

Abstract

Future space telescopes may be able to directly image $\sim 10 - 100$ planets with sizes
and orbits consistent with habitable surface conditions ("exo-Earth candidates" or EECs),
but observers will face difficulty in distinguishing these from the potentially hundreds of
non-habitable "false positives" that will also be detected. To maximize the efficiency of
follow-up observations, a prioritization scheme must be developed to determine which
planets are most likely to be EECs. In this paper, we present a Bayesian method for es-
timating the likelihood that any directly imaged extrasolar planet is a true exo-Earth can-
didate by interpreting the planet's apparent magnitude and separation in light of existing
exoplanet statistics. As a specific application of this general framework, we use published
estimates of the discovery yield of future space-based direct imaging mission concepts to
conduct "mock surveys" in which we compute the likelihood that each detected planet
is an EEC. We find that it will be difficult to determine which planets are EECs with
$> 50\%$ confidence using single-band photometry immediately upon their detection. The
best way to reduce this ambiguity would be to constrain the planet's orbit by revisiting the
system multiple times or through a radial velocity precursor survey. Astrometric or radial
velocity constraints on the planet's mass would offer a lesser benefit. Finally, we show
that a Bayesian approach to prioritizing targets would improve the follow-up efficiency
of a direct imaging survey versus a blind approach using the same data. For example, the
prioritized approach could reduce the amount of integration time required for the spectral
detection (or rejection) of water absorption in most EECs by a factor of two.

3.1 Introduction

One of the primary science goals driving the development of new space telescopes is the detection and characterization of Earth-like planets around nearby stars. Spectroscopy of the planets' reflected light spectra would reveal whether they are potentially habitable – i.e., liquid water could exist on their surfaces. The presence of biosignature gasses such as O_2, O_3, and CH_4 could be interpreted as evidence for life beyond the Earth (e.g., Seager et al., 2016; Fujii et al., 2018), although this interpretation would not be straightforward as many of these gasses can be produced abiotically (e.g., Catling et al., 2018; Meadows et al., 2018b).

Recently, final study reports have been published for two space telescope design concepts with a primary science goal of directly imaging and spectroscopically characterizing Earth-like planets around nearby stars. These are the Habitable Exoplanet Observatory (HabEx, Gaudi et al., 2020) and the Large UV/Optical/Infrared Surveyor (LUVOIR, The LUVOIR Team, 2019, hereafter L19). Following Stark et al. (2014), both reports provide estimates for the expected yield of planets across a range of sizes and insolations. The HabEx report predicts the detection and characterization of 12^{+18}_{-8} approximately Earth-sized planets in the habitable zone with a 4-meter aperture, while the LUVOIR report predicts 51^{+75}_{-33} with the 15-meter aperture "LUVOIR-A" design. These design concepts have been thoroughly investigated, and it is likely that the design of any future direct imaging space telescope would enable it to detect comparable numbers of potentially habitable planets with moderate S/N photometry.

However, these dozens of "exo-Earth candidates" would be detected amidst hundreds of planets with atmospheric compositions or equilibrium temperatures not conducive to Earth-like life, including planets outside of the habitable zone, large mini-Neptunes with thick H/He envelopes, or Mars-sized planets that have been stripped of their atmospheres. These non-habitable planets often demonstrate the same *observable* parameters (e.g., apparent magnitude and separation) as the exo-Earth candidates, but are far more common and may therefore cause a significant number of false positive detections. We demonstrate this problem in Figure 3.1. In fact, Guimond and Cowan (2018) show that separation- and

contrast-based selection criteria could suffer from a false discovery rate as high as 77% given just the detection data in a single band, or 47% if prior constraints on the orbit are available.

While the "false positive" planets would be interesting to characterize in their own right, spectroscopy across the full wavelength range could take *weeks* for the faintest targets, diverting time and resources from higher priority targets. To separate the potentially habitable and non-habitable planets, a survey could use multi-epoch broadband photometry to characterize their orbits and spectroscopic observations of H_2O absorption features to confirm the presence of water vapor in habitable zone targets. However, even within the habitable zone, there exist many planets that are too small or large to be habitable, and the identification of water absorption features will require a significant investment of time (Kawashima and Rugheimer, 2019). Additionally, warmer planets within the runaway greenhouse limit could exhibit water absorption features. In order to maximize the efficiency of an exo-Earth imaging mission, it is necessary to develop a method for identifying those planets that are most likely to be Earth analogs using the limited data that will be available upon detection.

We have previously advocated for a Bayesian approach to assessing the potential habitability of newly detected exoplanets (Apai et al., 2018a). The Bayesian approach allows one to probabilistically constrain the properties of the planet by leveraging knowledge from exoplanet statistics on planet radii, masses, and orbital properties. It also allows one to fold in predictions from theoretical models of planet formation and evolution as prior knowledge. As an example, in Bixel and Apai (2017) we used the Monte Carlo method to infer the likely composition of Proxima Cen b (Anglada-Escudé et al., 2016) in light of well-established statistical priors and the limited data available about the system. We found that it is $\sim 90\%$ likely that the planet is small and rocky as opposed to a "mini-Neptune". In this paper, we extend this approach to assess the likelihood that a directly imaged planet has an appropriate composition and orbit to be potentially habitable.

In Section 3.2 we review our Bayesian framework and give an example of how it can be applied to characterize directly imaged planets. In Section 3.3 we discuss the prior assumptions upon which our framework is based, and how they might be improved in

the coming decade. Using the planet yield estimates provided in LUVOIR final report as a baseline estimate for the yield of a hypothetical direct imaging mission, in Section 3.4 we conduct mock surveys where we detect each target and estimate the apparent likelihood that it has a potentially habitable composition and orbit. In Section 3.5 we discuss the results of these surveys, including what types of non-habitable planets would be mistaken for potentially habitable planets, and which additional data could help to resolve this ambiguity. Finally, we show that our approach to target prioritization could greatly enhance the efficiency of follow-up observations after all of the planets have been detected.

3.2 A Bayesian Framework for Classifying Directly Imaged Planets

3.2.1 Monte Carlo method

Here we review the Monte Carlo method for Bayesian inference. This method allows an observer to constrain the *unobservable* properties of a planet based on limited precision measurements of its observable data values by assuming some understanding of the prior distribution of intrinsic properties and their relationship to the observable data values.

Suppose planets can be described by some set of intrinsic properties $\boldsymbol{\theta}$ and some resulting set of observable data values $\mathbf{x}$ that can be calculated from $\boldsymbol{\theta}$. Given a prior probability distribution for the values of $\boldsymbol{\theta}$, $P(\boldsymbol{\theta})$, then the probability distribution for $\boldsymbol{\theta}$ *given* $\mathbf{x}$ can be calculated using Bayes' equation:

$$P(\boldsymbol{\theta}|\mathbf{x}) = \frac{P(\mathbf{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_\theta P(\mathbf{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})}$$

The left term is commonly referred to as the posterior distribution of $\boldsymbol{\theta}$, and $P(\mathbf{x}|\boldsymbol{\theta})$ as the likelihood function.

For most astrophysical applications there is no analytical solution to this equation, so it must be solved numerically. Under the Monte Carlo method, we use the prior probability distribution $P(\boldsymbol{\theta})$ to simulate a set of properties $\boldsymbol{\theta}_{\text{sim}}$, then calculate the simulated data values $\mathbf{x}_{\text{sim}}$ directly. Next, we accept or reject this simulated planet based on the value of its likelihood function for some observed set of data values $\mathbf{x}_{\text{obs}}$. Assuming the data

False positives for exo-Earth candidate detections

Size: ● too small   ● Earth-sized   ● too large
Orbit: ----- too close   ----- in habitable zone   ----- too far
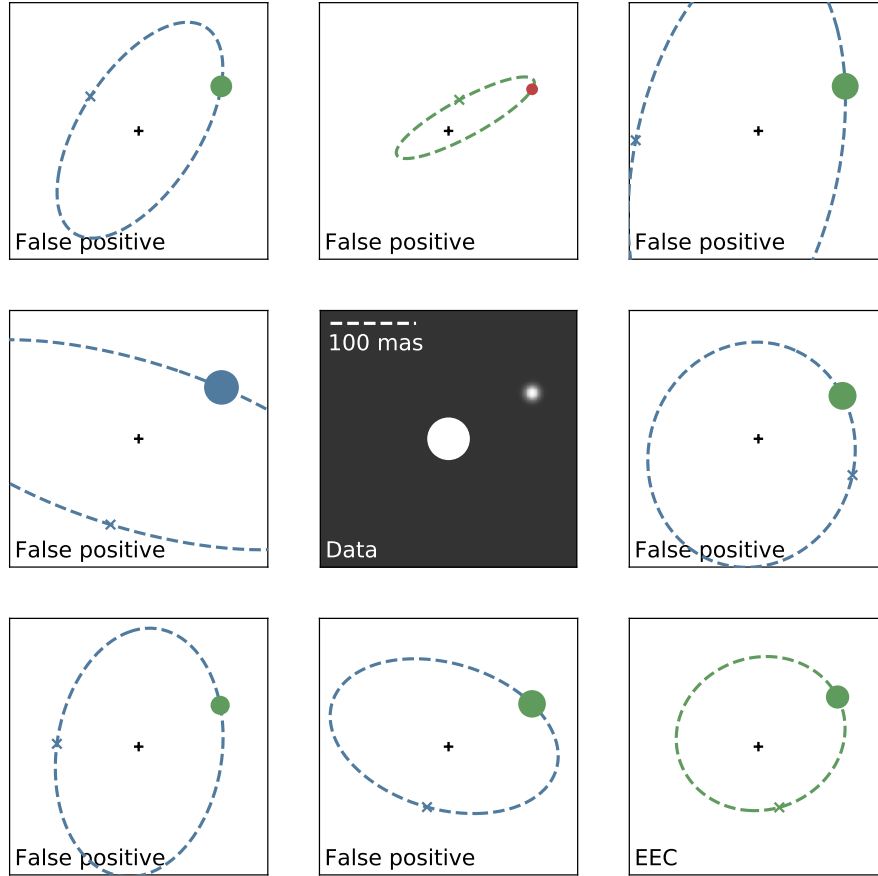
Figure 3.1: To illustrate the degeneracies that affect the interpretation of direct imaging data, we simulate the detection of a planet orbiting a Solar-type star at 15 parsecs (center panel), as well as several planets of varying sizes, orbits, and albedos that have a similar projected separation and magnitude (surrounding panels). It is not clear whether this data point represents a true exo-Earth candidate, or one of many potential false positives. The color and size of each circle represents the planet's potential radius; only green points are approximately Earth-sized ($\sim 0.8-1.6R_{\oplus}$). The color of the potential orbit represents its insolation; only green orbits are in the habitable zone. An 'x' marks the planet's closest approach to the observer.

values are drawn from independent, normal probability distributions with standard deviations (measurement uncertainties) σ_{obs}, the likelihood function is that of a multivariate Gaussian:

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{m} \exp((\mathbf{x}_{\text{obs},i} - \mathbf{x}_{\text{sim},i})^2 / 2\sigma_{\text{obs},i}^2)$$

If only an upper limit is available for a component of $\mathbf{x}_{\text{obs}}$ (e.g., magnitude), then the prior sample is first pruned of simulated members exceeding that limit.

This procedure is repeated in parallel for a large number of planets until a statistically sufficient number are accepted. The result of the likelihood-based selection is a sample of planets whose properties $\boldsymbol{\theta}_{\text{sim}}$ are distributed according to the posterior distribution. In other words, a histogram of the accepted values of $\boldsymbol{\theta}_{\text{sim}}$ represents the probability distribution for $\boldsymbol{\theta}_{\text{obs}}$, the properties of the observed planet.

Table 3.1.   List of Prior Assumptions Used to Build our Sample in Case 1

Parameter	Description of prior	Reference
Radius and period	*Kepler* occurrence rates	Mulders et al. (2018)
Class	"sub-terrestrial", "terrestrial", or "ice giant" based on radius, following Figure 3.4	Fulton et al. (2017); Zahnle and Catling (2017)
Mass	Empirical mass-radius relationship with intrinsic spread	Wolfgang et al. (2016)
Habitable zone boundaries	Planet mass-dependent LWHZ models (runaway and maximum greenhouse limits)	Kopparapu et al. (2014)
Albedo	Uniform from 0.2 to 0.7	
Eccentricity	Beta distribution ($\alpha = 0.867$, $\beta = 3.03$)	Kipping (2013b)
ω, Ω, M [†]	Uniform from 0 to 2π	
cos(i)	Uniform from -1 to 1	
Exo-Earth Candidates (EECs)	"terrestrial" class planets in the LWHZ	

Note. — References are given for each assumption. The additional cases in Table 3.2 may modify these assumptions to reflect new prior knowledge or data.

[†]Argument of pericenter; longitude of the ascending node; mean anomaly

3.2.2 Constructing the prior sample

We use the priors in Table 3.1 to construct the prior sample according to the algorithm visualized in Figure 3.2, and we discuss the prior assumptions in detail in Section 3.3. We consider several cases governing the amount of data available to the observer - these are listed in Table 3.2.

Before we simulate the properties of the directly imaged planet, we first consider the properties of its host star, which the observer will only know with finite precision. We represent this measurement uncertainty by drawing a unique stellar radius and mass for each system from normal distributions with σ = 3% and 7%. We assume nearly exact measurements of the distance and luminosity of the host star, and we discuss these assumptions in Section 3.3.1.

We generate the radius and period of the planet using *Kepler* occurrence rates as a prior probability distribution, then use the radius to classify the planet as too small ("sub-terrestrial"), too large ("ice giant"), or of the proper size to maintain a habitable atmosphere against stellar irradiation ("terrestrial"). This requires us to extrapolate to planets smaller ($R \lesssim 0.5 R_\oplus$) or with longer periods ($P \gtrsim 100$ days) than those readily available in the *Kepler* sample, as we discuss in Section 3.3.2. We calculate the mass from an empirical mass-radius relationship, where we assume some intrinsic variance due to stochastic planet formation histories and differences between the host stars of the planets on which these relationships are based. We draw eccentricities from a beta distribution and the remaining orbital elements are assumed to be isotropically distributed.

The liquid water habitable zone (hereafter LWHZ) is the range of orbital separations over which a broadly Earth-like planet could feasibly host liquid water on its surface. Kopparapu et al. (2014) find the zone's boundaries to be a function of the planet's mass and the star's effective temperature, hence we cannot infer a planet's membership to the LWHZ based solely on its insolation. We calculate the LWHZ boundaries and determine whether the planet lies in the runaway greenhouse, temperate, or maximum greenhouse regimes, interpolating between the discrete planet masses modeled by Kopparapu et al. (2014) (0.1, 1.0, and 5.0 $M_\oplus$), and taking the minimum or maximum mass values for

Table 3.2.   Different Cases Under Which we Conduct our Mock Surveys

Case	Description
1	Each planet has a monochromatic geometric albedo drawn uniformly from 0.2 to 0.7. The planet's monochromatic magnitude and separation angle are observed in a single epoch with a signal-to-noise ratio of 7 and centroid precision $\sigma_c = 3.8$ mas.
2	Prior radial velocity (RV) observations provide constraints of $\pm 10\%$ on the period, ± 5 cm/s on the radial velocity semi-amplitude, and $\pm 30°$ on the mean anomaly, in addition to the data from Case 1.
3	Simultaneous observations of a debris disk provide a $\pm 1°$ constraint on the inclination and longitude of the ascending node of the orbital plane. The planet's orbital elements may be further offset from these by ± 0.2 in Ω and $\cos(i)$.
4	Multiple epochs of direct imaging data permit constraints of $\pm 15°$ on the phase angle, $\pm 10\%$ on the semi-major axis, and ± 0.05 on the eccentricity..
5	Case 4 with an additional $\pm 0.1 \mu$as constraint on the semi-amplitude of the star's motion due to the planet.
6	We assume prior knowledge about the albedo distribution of small planets. Each planet's spectral albedo is determined by its class and location with respect to the habitable zone. Random monochromatic and polychromatic offsets are also introduced. The planet's magnitude in three bands and separation angle are observed in a single epoch with a wavelength-integrated signal-to-noise ratio of 7.
7	Combination of Cases 5 and 6: three-band photometry is available along with constraints on the planet's orbit and the star's astrometric motion from multi-epoch observations.

Note. — Each case represents different combinations of data that the observer might have.

Simulating the prior sample of planets

Priors

Properties

Spectral characterization → Host star properties

Occurrence rates → Period

Occurrence rates → Radius

Radius-based classification → Class

Mass-radius relationship → Mass

LWHZ model → Habitable zone boundaries

Isotropic priors → cos(i), Ω, ω, θ

Beta prior → Eccentricity

Uniform prior → Geometric albedo

Is it an exo-Earth candidate (EEC)?

Observables

Magnitude(s)

Apparent separation

K (cm/s) — Case 2

θ (µas) — Cases 5 & 7

Figure 3.2: Flowchart illustrating our algorithm for simulating the prior sample of planets under Case 1. The red boxes represent the intrinsic planet properties that we simulate, and the red arrows indicate how they are used in the calculation of other properties. The blue boxes represent the priors that we use to simulate the properties, and the green boxes are the observable values that can be compared to the data.

planets outside of this range. For planets with non-circular orbits we use the mean flux approximation to determine whether they are in the habitable zone, which Bolmont et al. (2016) find to be valid for planets with low or moderate eccentricities receiving Earth-like insolations.

Finally, we assign a geometric albedo to each planet using one of two methods for each of the cases in Table 3.2. In Cases 1-5, we assign a monochromatic albedo from a broad uniform prior with no dependence on the planet's class or orbital parameters. In Cases 6 and 7, we generate a spectrum for each planet by mapping it to a solar system analog with a comparable size and orbit. These two cases allow us to test the usefulness of color information for identifying potential exo-Earths. We dedicate more considerable discussion to the geometric albedos in Section 3.3.6.

3.2.3 Calculating the observable data values

Once the full assortment of planet properties has been simulated, we can proceed to calculate the observable data values to compare against those of a newly detected planet.

3.2.3.1 Apparent separation

The angular separation vector has two components, and can be calculated from the orbital elements and the distance to the system d. We adopt the same reference frame and notation as Murray and Correia (2010)[1], where $i = 90°$ is an "edge-on" inclination and the observer is at $z = \infty$, so the angular separation components orthogonal to the line of sight are:

$$s_x = (r/d)[\cos(\Omega)\cos(\omega+f) - \sin(\Omega)\sin(\omega+f)\cos(i)]$$

$$s_y = (r/d)[\sin(\Omega)\cos(\omega+f) + \cos(\Omega)\sin(\omega+f)\cos(i)]$$

In most cases, the position angle has no meaningfully defined zero point, so for simplicity we calculate only the net separation $|s| = \sqrt{s_x^2 + s_y^2}$. The exception is Case 4, where

---

[1]Figure 4 and Equations 53 & 54

the coeval detection of a debris disk is used to constrain the orbital plane of the planet; therefore the position angle is meaningfully defined, and we compute s_x and s_y separately.

3.2.3.2   Apparent magnitude(s)

Following Madhusudhan and Burrows (2012)[2], we model the planet as a Lambertian sphere, in which case the planet-to-star contrast ratio when observed at orbital phase α is:

$$\frac{L(\lambda)}{L_*(\lambda)} = A_g \left(\frac{R_P}{a}\right)^2 \left[\frac{\sin(\alpha) + (\pi - \alpha)\cos(\alpha)}{\pi}\right]$$

The phase angle is $\alpha = \mathrm{Cos}^{-1}[\sin(\omega + f)\sin(i)]$. We draw the geometric albedo from a prior distribution, and allow slightly super-Lambertian values ($A_g > 2/3$) as these are observed in some wavelength ranges in the solar system.

3.2.3.3   Radial velocity and astrometric semi-amplitudes and periods

The semi-amplitude of the star's periodic radial velocity variation, assuming no other perturbers and $M_P \ll M_*$, is:

$$K = (8.95\,\mathrm{cm/s})\frac{(M_P/M_\oplus)\sin(i)}{(M_*/M_\odot)^{1/2}(a/\mathrm{AU})^{1/2}(1-e^2)^{1/2}}$$

The semi-amplitude of the star's astrometric motion is:

$$\theta = (3.00\,\mu\mathrm{as})\frac{(M_P/M_\oplus)}{(M_*/M_\odot)}\frac{(a/\mathrm{AU})}{(d/\mathrm{pc})}$$

In both cases the period of the stellar motion (and therefore the planet's orbit) is also measurable. However the astrometric mass measurement requires that the system be observed for multiple epochs, in which case the planet's orbital period can be derived from its apparent motion about the host star.

---

[2]Figure 1 and Equations 4 & 33

3.2.4 Calculating the posterior probability distributions

We generate a posterior sample of simulated planets following the scheme in Section 3.2.1, where $\boldsymbol{\theta}_{\mathrm{sim}}$ are the simulated properties from Section 3.2.2, $\mathbf{x}_{\mathrm{sim}}$ are the simulated data values from Section 3.2.3, and $\mathbf{x}_{\mathrm{obs}}$ are the data values for the observed planet. This allows us to calculate a posterior probability distribution for each component of $\boldsymbol{\theta}_{\mathrm{obs}}$, the physical properties of the observed planet. Based on just the planet's apparent separation and magnitude, we are therefore able to place informative constraints on its semi-major axis, radius, and mass.

The posterior probability in which we are most interested is the probability that the planet is potentially habitable - i.e., a "terrestrial"-class planet $(0.8 \lesssim R_P \lesssim 1.6)$ within the LWHZ. We designate these planets as "exo-Earth candidates" (EECs), following the terminology of The LUVOIR Team (2019) and Gaudi et al. (2020), albeit including a slightly different range of sizes. Planets outside of this range we designate as "false positives", with the *false positive probability* being the likelihood that the observed planet - as judged based solely on its observed data values - is a false positive instead of an EEC.

3.2.5 Example: an exo-Earth candidate around a Solar twin

As a demonstration of our method, we use the procedure in Section 3.2.2 to generate an Earth-sized planet in the center of the LWHZ of a Solar-type star at 15 parsecs - an ideal exo-Earth candidate. We simulate the detection of this planet by assuming a S/N = 7 measurement of its monochromatic magnitude, and a ± 3.8 mas measurement of its centroid.

Acting as the observer - who has no prior knowledge about the planet's true size and orbit - we use the procedure outlined above to estimate the likelihood that that the planet is an exo-Earth candidate based on its observed magnitude and separation. The results of our analysis are summarized in Figure 3.3. We can confidently say that the planet is at least 1 AU from its star, and is unlikely to be farther than 3 AU - however, we cannot constrain its orbit to the habitable zone with certainty. We can also tell that the planet is almost certainly larger than $0.5 R_\oplus$ and smaller than $3 R_\oplus$ - but this range includes planets
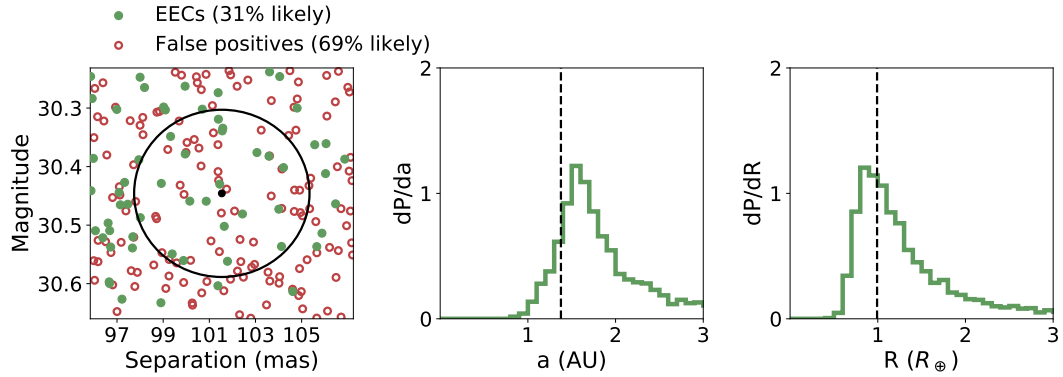
Figure 3.3: We simulate the detection of an "ideal" exo-Earth candidate from 15 pc. (Left) The separation- and magnitude- phase space populated by a host of simulated EECs (green) and non-EECs (red). The data point with uncertainties is marked in black. (Middle/right) Posterior distributions for the planet's semi-major axis and radius, taking into account *Kepler* statistics and other priors. The true values are marked with dashed lines; this planet appears to the observer to have a wider orbit than it actually does.

that are too small (sub-terrestrial) or too large (sub-Neptune) to be EECs.

Most notably, we determine that it is only 31% likely that this planet is an EEC, as the majority of simulated planets that have a similar separation and magnitude are not habitable. In this specific example, even though the planet appears 70% likely to have a *size* consistent with habitability, it also appears 60% likely to orbit beyond the maximum greenhouse limit. This example shows that while it will be difficult to discriminate between true EECs and false positives given just the data available on detection, it will still be possible to place meaningful probabilistic constraints on the planet's properties.

3.3 Prior assumptions

In this section, we discuss our priors in more detail by reviewing the relevant literature and discussing how they may be improved upon by future observations and modeling efforts.

3.3.1 Stellar properties

It is likely that much effort will be dedicated to characterizing the stellar targets of a direct imaging mission in advance of its launch. Still, the stellar properties will only be constrained with finite precision - potentially several percent - so it is important that our prior sample includes host stars spanning the range defined by the relevant uncertainties.

Gaia DR2 (Gaia Collaboration, 2018) has already provided high precision ($10-100$ μas) parallax measurements for nearby F-M spectral type stars, so we treat this uncertainty as negligible. Optical/IR interferometry has allowed for the measurement of stellar radii to $\sim 3\%$ for targets at $\sim 10-100$ pc (e.g. Ligi et al., 2016). Masses are more difficult to measure, so stellar atmosphere models are often used - as an example, Sharma et al. (2018) constrain model-dependent masses for more than 10,000 stars using high resolution spectroscopy, with a median precision of 7%. Following these examples, we draw the radius and mass of the host star from normal distributions with widths of 3% and 7%, respectively.

3.3.2 Radius and period

Kepler allowed for the precise calculation of planet occurrence for planets with periods shorter than 100 days. Mulders et al. (2018) find that the Kepler occurrence rates are well-described by independent broken power laws in both radius and period:

$$\frac{dN_{\mathrm{pl}}}{d\log P\, d\log R} \propto f_R(R) f_P(P)$$

where

$$f_P(P) = (P/P_{\mathrm{break}})^{a_P} \quad (P < P_{\mathrm{break}})$$
$$= (P/P_{\mathrm{break}})^{b_P} \quad (P \geq P_{\mathrm{break}})$$

$$f_R(R) = (R/R_{\mathrm{break}})^{a_R} \quad (R < R_{\mathrm{break}})$$
$$= (R/R_{\mathrm{break}})^{b_R} \quad (R \geq R_{\mathrm{break}})$$

The best-fit parameters for this model are $(P_{\mathrm{break}}, a_P, b_P) = (12, 1.5, 0.3)$ and $(R_{\mathrm{break}}, a_R, b_R) = (3.3, -0.5, -6)$. The number of planets per system, N_{pl}, is found to be ~ 5. Multiplicity will be an important factor for direct imaging surveys, as it could confuse the interpretation of the data or allow for simultaneous follow-up observations of multiple planets - but the topic is outside of the scope of this work. Here, we treat each detected planet independently, and normalize the above power laws so that $N_{\mathrm{pl}} = 1$.

To properly simulate the abundance of planets in the habitable zone of F, G, and K spectral type stars - as well as smaller or cooler planets that might be mistaken for them - requires us to extrapolate *Kepler* occurrence rates beyond the range of parameters within which they are well-understood ($R \gtrsim 0.5 R_{\oplus}$, $P \lesssim 100$ days). This extrapolation could be problematic; for example, the results of Chen and Rogers (2016) suggest that planets on wider orbits can maintain thick volatile envelopes better against hydrodynamical escape, so we might find an over-abundance of large planets on wide orbits. We can gain some insight by studying the dependence of planet radii on insolation around low-mass stars, but these results cannot necessarily be extrapolated to Solar-mass regimes.

Furthermore, while *Kepler* was generally not sensitive to planets smaller than $0.5 R_{\oplus}$, it is likely that some such planets will be found by direct imaging missions and could masquerade as exo-Earth candidates. It is therefore necessary that we extrapolate the power law of Mulders et al. (2018) down to $0.1 R_{\oplus}$ to ensure that the potentially large number of Mercury-sized objects are represented in our simulations. However, since our cutoff for exo-Earth candidates is $R \approx 0.8 R_{\oplus}$, planets smaller than $0.5 R_{\oplus}$ are less likely to be mistaken for EECs, so this extrapolation should not substantially influence our results.

3.3.3 Planet classes

We employ a radius-based classification scheme to separate potentially habitable planets from those that are too small or too large to be habitable. This classification is motivated by two physical considerations affecting whether a planet can maintain a habitable atmosphere against irradiation over several Gyr.

Empirical evidence suggests a change in planet compositions between $1.5 - 2 R_{\oplus}$. Multiple authors find evidence for a split in planet densities in this range, with planets

larger than $\sim 1.5\,R_\oplus$ mostly having densities much lower than the Earth's (e.g., Weiss and Marcy, 2014; Rogers, 2015; Chen and Kipping, 2017). Fulton et al. (2017) find a relative lack of *Kepler* planets with $R \sim 1.75\,R_\oplus$ compared to smaller or larger radii; this "photoevaporation valley" was predicted by several authors who show that smaller planets would lose thick envelopes due to hydrodynamic escape (e.g., Owen and Wu, 2013). We interpret both results as evidence that planets larger than $\sim 1.4 - 1.7\,R_\oplus$ have compositions more comparable to Neptune than the Earth, and are therefore not habitable in the traditional sense.

Very small planets will also have trouble maintaining even small and dense atmospheres against Earth-like insolations. Zahnle and Catling (2017) find that a simple power law relationship between a body's escape velocity and its effective insolation ($I \propto v_{\mathrm{esc}}^4$) can predict whether planets in the solar system (and some exoplanets) have atmospheres. According to this relation, a planet with the same insolation and density as the Earth would need to be larger than $0.8\,R_\oplus$ to maintain a habitable atmosphere. However, we note that the Earth lies towards the inner edge of the LWHZ as calculated by Kopparapu et al. (2014); it is possible that smaller planets could maintain Earth-like atmospheres further out.

Taking both of these considerations into account, we assign one of three classes to each planet based on its radius: "sub-terrestrial" planets that are too small to be habitable, "terrestrial" planets that could have Earth-like atmospheres, and "ice giant" planets that are too large. There is likely some overlap between these categories; for example, planets slightly larger than $1.4\,R_\oplus$ or smaller than $0.8\,R_\oplus$ might still have an Earth-like atmosphere. To simulate this overlap we probabilistically assign each planet's class from its radius using the probabilities in Figure 3.4. The "terrestrial" class includes all planets with $0.8 < R < 1.4\,R_\oplus$ and a fraction of planets with $0.5 < R < 0.8\,R_\oplus$ or $1.4 < R < 1.7\,R_\oplus$.

3.3.4 Mass

To calculate each planet's mass, we rely on the empirical mass-radius relationships of Wolfgang et al. (2016)[3] (hereafter W16), which are calculated for smaller ($< 1.6\,R_\oplus$) and

---

[3]Equation 2 and Table 1
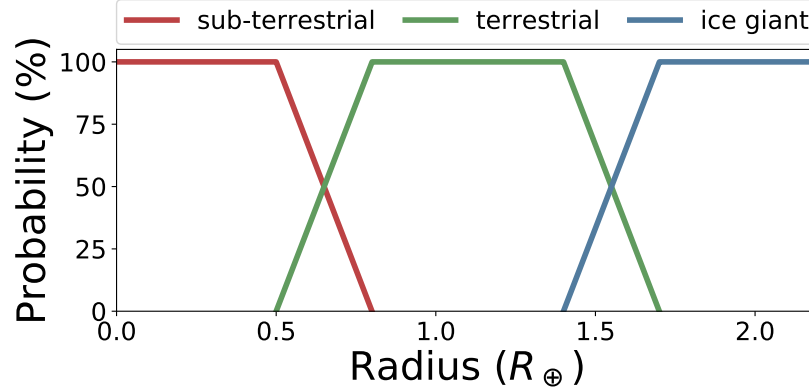
Figure 3.4: Our probabilistic scheme for classifying planets based on their radii. "Sub-terrestrial" planets are so small that they will lose their atmospheres to thermal escape under LWHZ levels of irradiation. "Ice giants" are so massive that they will form and maintain thick volatile envelopes. Only the "terrestrial" planets are neither too small nor too large to maintain a habitable atmosphere against irradiation.

larger ($< 4R_\oplus$) planets, reflecting the bimodal split in planet compositions. They do not treat mass as a deterministic function of radius, but rather model a *distribution* of masses for each radius to capture the intrinsic variability in planet compositions.

We draw planet masses from truncated normal distributions defined by mean μ and variance σ^2, with minimum values of $0.01\,\mu$ and maximum values of $M_{\text{pure Fe}}$ - the mass of a pure iron composition as defined in W16. The parameters of the distributions are:

$$
\mu, \sigma = \begin{cases}
2.7\,M_\oplus\,(R/R_\oplus)^{1.3}\,,\ 1.9\,M_\oplus & (R \geq 1.6\,R_\oplus) \\
1.4\,M_\oplus\,(R/R_\oplus)^{2.3}\,,\ 0.3\mu & (0.8 < R < 1.6\,R_\oplus) \\
1.0\,M_\oplus\,(R/R_\oplus)^{3.0}\,,\ 0.3\mu & (R \leq 0.8\,R_\oplus)
\end{cases}
$$

These are the values of μ and σ fitted by W16, with a few caveats:

1. Their data did not allow the authors to determine σ_M for the smaller planets; here, we arbitrarily choose $\sigma = 0.3\mu$ (i.e., a 30% spread in density).

2. Only a few uncertain data points and upper limits were available for planets with $R < 0.8\,R_\oplus$, so we instead assume approximately Earth-like densities.

3. The large radius relationship was fitted for all planets with $R < 4R_\oplus$, not just the ice giants; however, most of the precise data points nevertheless had $R > 1.6R_\oplus$.

4. We model a few planets as large as $10R_\oplus$, but this mass-radius relationship is likely not valid beyond $4R_\oplus$; indeed, it underestimates the mass of Jupiter (~ 11 $R_\oplus$) by a factor of five. Ultimately, the overlap in masses and magnitudes between Earth-sized and Jupiter-sized planets is negligible when considering potential false positives for exo-Earth candidate detections.

Finally, we expect that the empirical mass-radius relationship will be improved upon in coming years by the discovery of transiting rocky planets around low-mass stars by TESS (Ricker et al., 2014), precision radius measurements from CHEOPS (Broeg et al., 2013), and mass measurements through TTV or RV. By the time an exo-Earth direct imaging mission begins, observers should have a better understanding of the relationship between a planet's size, composition, and mass with which to interpret radial velocity or astrometric mass measurements.

3.3.5   Eccentricity

Planets in the solar system tend to have eccentricities smaller than 0.1, but multiple authors find evidence for a wider distribution of exoplanet eccentricities in both transit (Kane et al., 2012) and radial velocity data (Kipping, 2013b). Kipping (2013b) determine that the eccentricity distribution of several radial velocity detected exoplanets is well-described by a beta function, with $\alpha = 0.867$ and $\beta = 3.03$ - in which case $> 50\%$ of planets have $e > 0.1$.

Eccentricity will have the effect of confusing the determination of a planet's orbit from a single epoch of imaging data, as a wider range of eccentric orbits could be consistent with the observed separation. Furthermore, planets which orbit near the inner or outer edge of the LWHZ may spend a fraction of their orbit outside of the zone. To determine which of these planets are EECs we use the mean flux approximation - assuming that a planet is habitable if the average insolation of its orbit is the same as a circular orbit in the habitable zone; Bolmont et al. (2016) find this to be an adequate approximation for

planets with low or modest eccentricities receiving a mean flux equal to the Earth's, while highly eccentric planets tend to freeze out. However, it is possible that this approximation is not valid for eccentric planets receiving a lower mean flux.

We use the beta distribution of Kipping (2013b) to draw planet eccentricities and compute the projected separation accordingly, though we truncate the distribution beyond $e > 0.8$ (0.5% of planets) due to the additional computing time required to solve Kepler's equation for highly eccentric orbits.

We note that multi-planet systems such as our own tend to have more circularized orbits (Van Eylen and Albrecht, 2015). In principle, if a planet is detected with a companion then it is *a priori* less likely to have an eccentric orbit, and it should be easier to determine whether the planet is an EEC. In the scope of this paper, however, we treat all planets as the only member of their system.

3.3.6 Albedo

We consider two different prescriptions for simulating the geometric albedo: in Cases 1-5 we assume a monochromatic albedo drawn from a broad uniform distribution, while in Cases 6 and 7 we assume a spectral albedo model that depends on the planet's class and position with respect to the LWHZ.

Constraining the actual distribution of planet surface and atmospheric properties is one of the goals of future imaging missions, so it might seem backwards to interpret these observations by assuming the underlying distribution of geometric albedos as a prior. Nevertheless, such an assumption is necessary in order to infer a planet's size from photometric data.

A uniform prior represents a conservative approach to the problem. We note that most planets in the solar system have geometric albedos ranging from 0.2 to 0.7 across the UV to NIR wavelength range, with the exception of Mercury's very low albedo. Therefore for most cases we draw a monochromatic geometric albedo for each planet uniformly from 0.2 to 0.7.

However, the proposed designs of direct imaging missions allow for the simultaneous observations of a planet in 2-3 photometric bands, in which case color information would

be available with no additional overhead. An efficient characterization strategy would make use of this color information to discriminate between EECs and their false positives, but to do so the observer must assume some prior knowledge about the diversity of planetary atmospheres and surfaces. As an example, we assume that all planets approximately reflect one of four solar system analogs - Earth, Venus, Mars, or Neptune - based on their radius-based classification and orbit as described in Table 3.3.

We accessed model spectra through the Virtual Planetary Laboratory[4] for the Earth (Robinson et al., 2011, scaled from quadrature), Venus dayside, and Mars (no publications listed). For Neptune we use the planet's observed geometric albedo as provided by Madden and Kaltenegger (2018) from 450-2500 nm, and set $A_g = 0.6$ from 300-450 nm (e.g., Mallama et al., 2017).

To simulate both model uncertainty and physical diversity among these solar system analogs, we allow the spectral albedos to vary by ± 0.1 monochromatically and ± 0.05 in each photometric bandpass. Additionally, we enforce a lower limit of 0.001 - to ensure that none of our planets are perfect blackbodies - and an upper limit of 0.7, slightly more than the upper limit for the geometric albedo of a Lambertian sphere. The range of spectral models for each category of planet is demonstrated in Figure 3.5.

Finally, we consider the bandpasses in which the planets are observed. For reference, LUVOIR's proposed coronagraphic instrument would be able to observe simultaneously in 10% bandpasses of each of its three channels (L19). We choose wavelength ranges near the centers of each channel: 335-390 nm, 715-830 nm, and 1390-1610 nm.

It is likely that a wide variety of terrestrial planets exist; indeed, all four of the terrestrial planets in the solar system are distinct in surface reflectance and atmospheric absorption. Both of the ice giants, however, have similar albedo distributions. This suggests that some understanding of the albedo distribution of ice giant analogues can be attained during the coming decade. Modeling efforts to understand the composition and appearance of sub-Neptune type planets are already underway (e.g., Hu et al., 2015). New observatories such as JWST, WFIRST, and ELTs could provide observational tests of these models - the first through eclipse and transit spectroscopy of warm and hot Neptunes, and the

---

[4]http://depts.washington.edu/naivpl/content/vpl-spectral-explorer

Table 3.3.   List of the Solar System Analogs Used to Simulate Each Planet's Spectrum
Under Cases 6 and 7

Class	Location	Assumed model
terrestrial	in LWHZ	Earth
sub-terrestrial	anywhere	Mars
terrestrial	exterior to LWHZ	Mars
terrestrial	interior to LWHZ	Venus
ice giant	anywhere	Neptune

Note. — The assumed model for the planet's spectrum depends on its class/size and location with respect to the liquid water habitable zone, as defined by the runaway and maximum greenhouse limits.

latter two through direct imaging of ice giants at wide separations.

3.4   Mock surveys: methodology

The LUVOIR team has released a final report (L19) that include estimates for the number of planets that LUVOIR could detect as a function of planet radius, insolation, and host spectral type. In this section, we produce a candidate sample based on these estimates, characterize each planet therein, and report on the efficiency of each strategy outlined in Table 3.2 for properly identifying exo-Earth candidates. While we rely on the yield estimates and stellar targets of L19 for our mock survey, we use these only as baseline estimates for the yield of a generic, hypothetical direct imaging survey; we do not attempt to reproduce the results of the report or to assess the mission design.
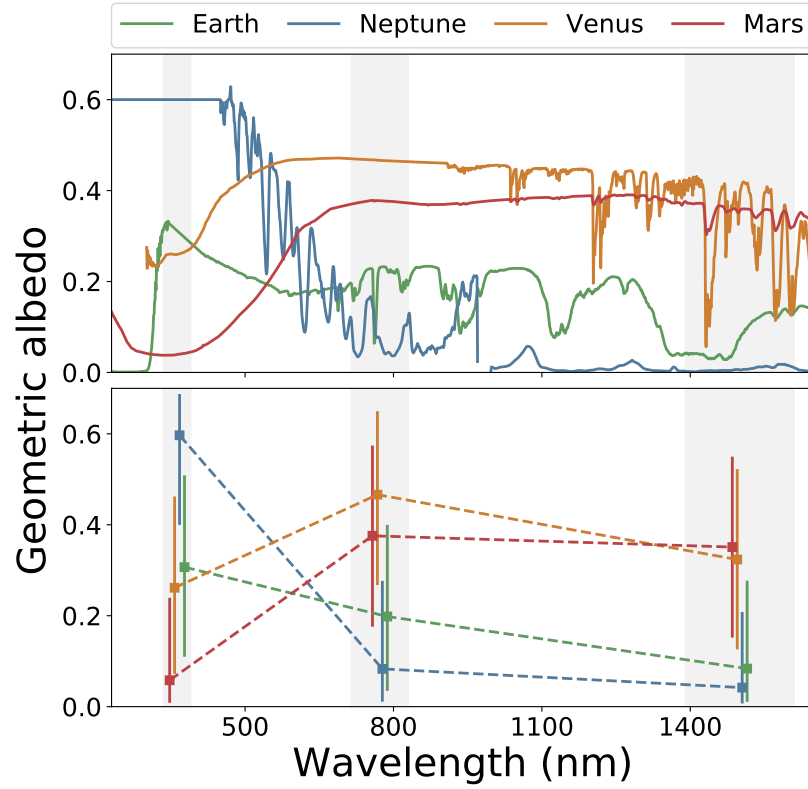
Figure 3.5: (Top) The set of base models that we use to simulate geometric albedos under Cases 6 and 7 only, according to the scheme in Table 3.3. (Bottom) The actual range of simulated spectra for each type of planet as observed in three photometric bandpass (with wavelength offsets for visibility). We apply moderate differences to each planet's spectrum to simulate the underlying physical diversity, but broad differences between the four groups can still be seen.

3.4.1   Generating the candidate list

To properly generate a list of planets that could be detected by a coronagraphic imaging mission requires a careful treatment of the instrument design, survey strategy, and additional sources of noise (e.g., dust) that is outside of the scope of this work. L19 have performed such an analysis (following the methodology laid out by Stark et al. (2014)) so we use their results to generate our candidate list[5].

We acquired one of the simulated samples of host stars upon which the LUVOIR-A yield estimates are based, including masses, luminosities, and distances for 287 stars (C. Stark, private correspondence). To assign a radius to each star we use a simple scaling relation (Hansen et al., 2004):

$$(R_*/R_\odot) = (M_*/M_\odot)^{0.8}$$

While this simulated list will not be the final target list of LUVOIR-A, it generally represents the diversity of host star properties that such a survey would encounter. Next, we use the method outlined in Section 3.2.2 to generate a large sample of planets around these stars, and draw from that sample at random until the yield estimates for LUVOIR-A have been satisfied for each bin in spectral type, planet size, and insolation.

To ensure that the planets we simulate are detectable, we also enforce the same separation and brightness limits as L19. Namely, we only include targets that are brighter than a planet-to-star contrast ratio of 2.5×10^{-11} and that are detected between 24–440 mas - approximately the working angles of the proposed LUVOIR-A coronagraph at 500 nm.

The yield estimates of the LUVOIR-A architecture project the discovery of ~ 450 planets, ~ 50 of which would be exo-Earth candidates. For our purposes, the actual number of planets is not relevant - only their relative abundance by size, distance, host spectral type, etc. - so we improve the accuracy of our results by inflating the yield estimates unilaterally by a factor of fifty.

---

[5]see Figures 3.1 & 3.13 and Tables 8.7 for details on the target list, yield estimates, and instrument parameters

3.4.2 Survey cases

We run our mock surveys under seven cases representing the different data that could available to the observer. A brief description of each can be found in Table 3.2.

3.4.2.1 Case 1: Detection data only

Under Case 1, the planet's existence is entirely unknown before its direct imaging detection, and the observer only measures its apparent separation and monochromatic magnitude. Following L19, we assume a signal-to-noise ratio of 7. If photon noise is dominant, the uncertainty on the planet's centroid position is described by:

$$\sigma_c = \frac{\text{FWHM}}{\text{SNR}}$$

Assuming an effective wavelength of 500 nm, then $\sigma_c = 3.8$, 1.9, and 1.0 mas for the 4-meter HabEx and 8- or 15-meter LUVOIR architectures. This uncertainty will be further affected by the pointing stability of the telescope during observations and the wavelengths at which the planet is observed; to be conservative, we choose $\sigma_c = 3.8$ mas. Finally, we simulate the planet's detection by re-drawing its magnitude and apparent separation from normal distributions with widths defined by these uncertainties.

3.4.2.2 Case 2: Additional radial velocity detection

It has been emphasized that a radial velocity search for nearby Earth twins would be an important precursor to a space-based direct imaging mission (Dressing et al., 2019), but the detection of Earth analogs around Solar twins is beyond the reach of current instrumentation. Doing so would require a significant investment of time on major observing resources and new methods to correct for systematic sources of noise such as stellar jitter (Plavchan et al., 2015).

To investigate the potential benefits of a precursor radial velocity search for interpreting planet detections, in Case 2 we simulate the direct imaging detection of a planet (Case 1) along with a measurement of its orbital parameters and radial velocity semi-amplitude.

We set a baseline uncertainty of 5 cm/s on the measurement of the radial velocity semi-amplitude K - this value is chosen to be smaller than the value for an Earth twin (~ 10 cm/s), but not negligible in comparison. Additionally we allow a conservative 10% uncertainty on the measurement of the orbital period, and a loose $\pm 30°$ constraint on the planet's mean anomaly.

Several planets - including small planets in the LWHZ that could masquerade as exo-Earths candidates - will have $K < 5$ cm/s. In these cases we assume the planet is undetected, so no constraints on its orbit are available. An upper limit of 10 cm/s is enforced so that the non-detection also carries useful information about the planet's size.

3.4.2.3 Case 3: Constraining the orbital plane using debris disks measurements

It is expected that several nearby systems contain exozodiacal dust disks near their habitable zones (Ertel et al., 2018). This dust may be a significant source of background and confusion noise for future direct imaging surveys and could be a key driver of aperture size, with larger apertures collecting less background light within their resolution element (Roberge et al., 2012). However, these disks could also carry useful information about the orientation of the system that could help to establish a newly detected planet's orbit, and would require no additional resources to observe.

Observations of larger protoplanetary disks have yielded tight (better than 1°) constraints on the disk inclination and orientation (e.g., HL Tau, ALMA Partnership et al., 2015). While the orientation of exozodiacal disks may be harder to constrain if they are faint or exceed the outer working angle of the coronagraph - and while not all systems may have substantial disks at all - we here consider the "optimistic" case where every system has a disk, and the orientation parameters ($\cos(i)$ and Ω) can be constrained to $\pm 1\%$.

In principle, if the orientation of the debris disk can be tightly constrained and the planet shares *exactly* its orbital plane with a circular orbit, then a single precise measurement of the planet's apparent separation could be sufficient to determine its phase and semi-major axis. However, the orbital inclinations of solar system planets deviate by up to 5° from the Solar spin axis - and larger misalignments could be possible in other

systems - while several of our simulated planets have eccentric orbits.

We can still infer some information about the planet's orbit given the orientation of the disk and some prior knowledge about how misaligned planetary systems tend to be. In Case 3, we use the orientation of a contemporaneously detected disk as a prior constraint on the mean orbital plane of the system, but allow for a difference of ±0.2 between $\cos(i)$ and the longitude of the ascending node (Ω) of the two components. We then interpret the observations using the same data as in Case 1, but treating the two dimensions of the separation vector separately.

3.4.2.4 Cases 4 and 5: Multiple revisits and astrometric mass measurements

Constraining the planet's orbital parameters will require multiple revisits spaced over the orbital period, so revisits will likely be folded into the observing strategy of future direct imaging missions. However, even within the habitable zone there will be numerous potential false positives for EECs, a fact that may limit the practical benefit of revisiting every system.

Guimond and Cowan (2019) have shown that ~ 3 revisits will be sufficient to constrain the orbital parameters with better than 10% precision. In Case 4 we assume that the system has been revisited enough times for the planet's orbital parameters to be constrained with comparable precision. Specifically, we assume that the semi-major axis is measured to $\pm10\%$, the eccentricity to ±0.05, and the orbital phase to $\pm15°$.

An ancillary benefit of revisiting targets would be the ability to measure the star's astrometric motion about the system's center of mass. Measuring the astrometric semi-amplitude θ would allow observers to determine the planet's mass and potentially to identify which planets are too small or large to be habitable. In Case 5 we assume that, in addition to the constraints provided in Case 4, the observer can measure θ with a precision of ±0.1 μas. This is the targeted astrometric precision of the High Definition Imager (HDI) instrument with the LUVOIR-A aperture (L19), and approximately $1/2$ of the amplitude induced by an Earth twin at a distance of 15 parsecs. If $\theta < 0.1$ μas, an upper limit of 0.2 μas is enforced instead.

3.4.2.5 Case 6: Color measurements

Future direct imaging missions could be able to observe simultaneously in multiple band-passes, in which case measurements of the planets' colors would be available as soon as they are detected. Color information could be used for preliminary planet classification; for example, Batalha et al. (2018) predict that it will be possible to differentiate between cloudy and cloud-free Jovian planets with reflected light imaging in three filters using WFIRST or ELTs.

However, color information is only useful for inferring the planet's properties if we make a prior assumption about the spectral albedos of extrasolar planets. Some observational constraints are currently available through eclipse observations of close-in giant planets, but for smaller planets we can only rely on planets in the solar system and theoretical models of the surfaces and atmospheres of known transiting and RV-detected planets. It is therefore worthwhile to determine what effect an approximate prior understanding of spectral albedos would have on the interpretation of direct imaging data.

In Case 6 we simulate planets with spectral albedos reflective of (though not identical to) solar system analogues; our detailed assumptions are described in Section 3.3.6. Each planet is observed in a 10% bandpass at the center of three wavelength channels, with a signal-to-noise ratio weighted by the square root of the bandpass-integrated flux (i.e., photon noise). The signal to noise integrated across all three bandpasses is 7, as in Case 1. If S/N < 2 in a given bandpass (typically in the UV or infrared), then a 2σ upper limit is enforced instead.

3.4.2.6 Case 7: Maximum information

In the final case we consider all of the information that a larger telescope would be able to acquire on a target after several revisits. These include measurements of the orbital parameters and astrometric semi-amplitude as well as brightness measurements in three bands. We assume the same measurement precision and sensitivity limits as in Cases 4 and 5, and we interpret the color information following the method of Case 6.

3.4.3 Classifying the targets

Once we have constructed a sample of targets and simulated their detection, we classify them using the inference framework described in Section 3.2. This yields for each planet the likelihood, according to the observer, that the planet is an EEC or a false positive. In Figure 3.6 we plot the likelihood that the planet is an EEC for each simulated EEC in the sample. Since the observer does not know *a priori* that these planets are EECs, we see that they cannot make a confident identification upon the planets' initial detections, but given additional data or multiple revisits they can identify several EECs with confidence.

We can break down the false positive probability by size and orbit, determining for each observed EEC the inferred likelihood that it is, for example, a sub-Neptune on an orbit exterior to the habitable zone. In Figure 3.7 we plot this probability for each combination of class and orbit averaged over the sample of observed EECs. This plot illuminates the key sources of ambiguity in classifying EECs. For example, we see that it is difficult to distinguish between an EEC and a sub-terrestrial planet with a temperate orbit, or a planet that has the proper size but lies just interior or exterior to the habitable zone.

3.5 Mock surveys: results

3.5.1 Which planets are "false positives"?

Figure 3.7 demonstrates that true EECs share the observable parameter space with a wide range of planets both within and outside of the LWHZ. These can be broadly separated into: (i) planets that are too small or large to be habitable, but are yet in the habitable zone, (ii) planets that are of the proper size to be habitable, but are not in the habitable zone and (iii) planets that are both of the wrong size and not in the habitable zone. Each of these categories are approximately equal in their potential to masquerade as EECs.

As we demonstrate in Figure 3.6, the observer will typically be unable to distinguish between true EECs and their many potential false positives given just the data necessary for the planet's detection. Even when a planet that is *in fact* an EEC is detected, the observer will only be able to make this determination with $< 50\%$ confidence.

How confidently can we identify exo-Earth candidates in mock observations?
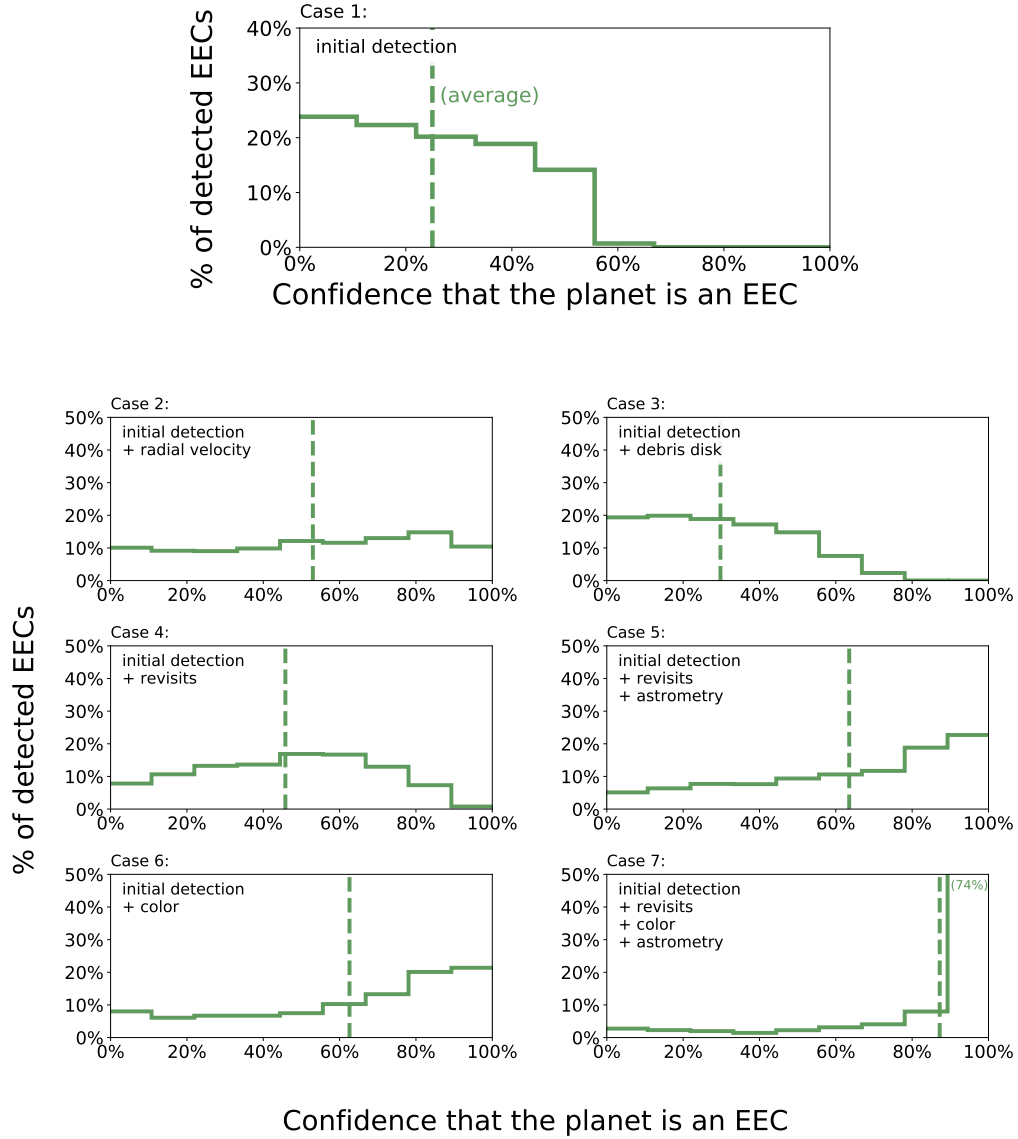
Figure 3.6: We simulate the detection of ~ 2500 EECs around nearby stars, then infer their properties from the mock data under each of the cases in Table 3.2. Above we plot the probability - as inferred by an uninformed observer - that each detected planet is an EEC, as well as the average value (dashed line). In the ideal case, this value would be 100% for all EEC targets, but typically it is smaller because of the limited data available to the observer. With additional data (Cases 2–7), the observer can be more confident that the detected planet is an EEC.

What kinds of planets could be mistaken for exo-Earth candidates?
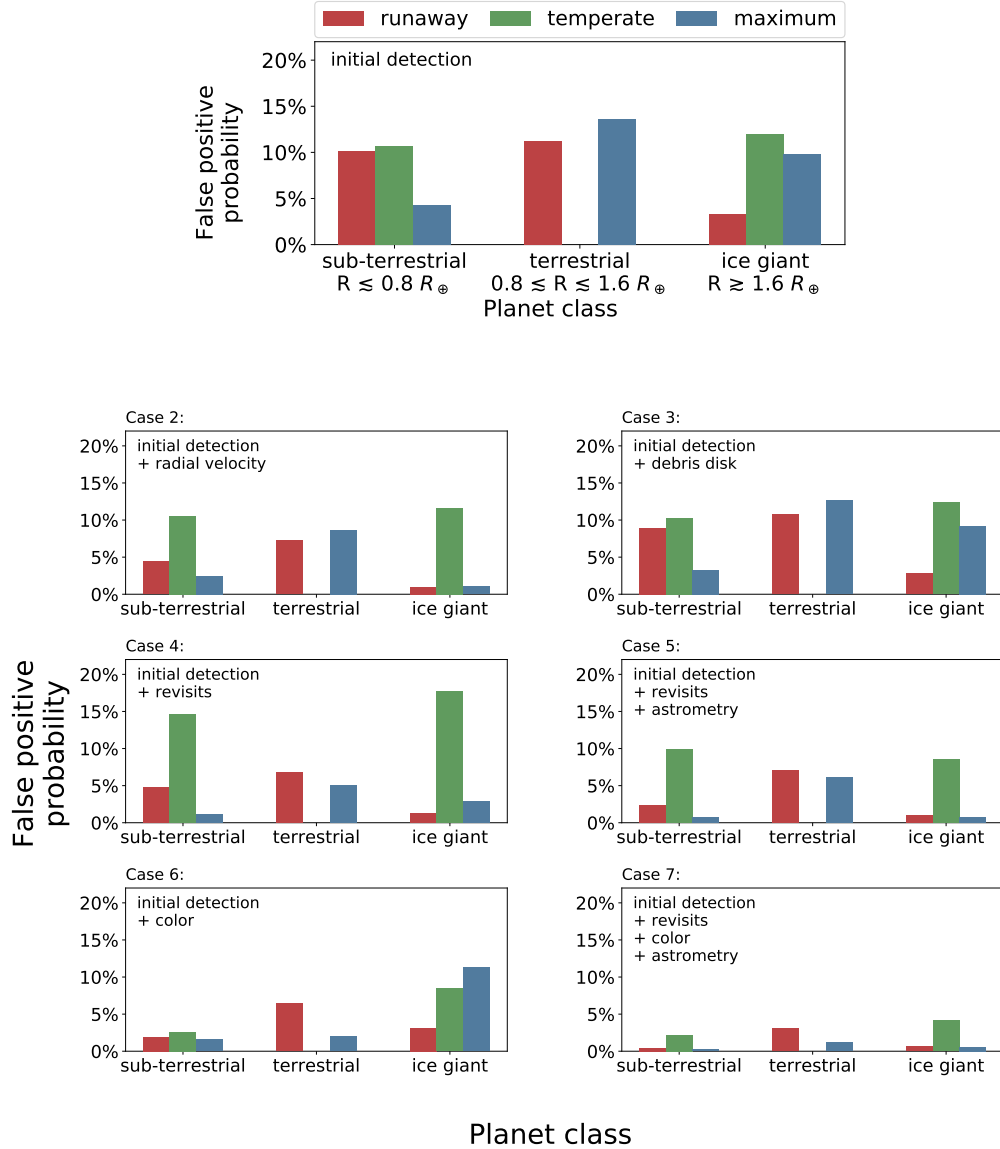
Figure 3.7: We simulate the detection of ~ 2500 EECs around nearby stars, then infer their properties from the mock data under each of the cases in Table 3.2. Above we plot the sample-averaged probability - as inferred by an uninformed observer - that the planet is instead a false positive with a non-habitable class or orbit. Additional data will suppress the false positive probability - for example, fitting the orbit with finite precision will reduce the likelihood that the planet is outside of the habitable zone.

3.5.2 Do constraints on the orbit help to identify EECs?

Since most of the "false positives" are planets outside of the habitable zone, it stands to reason that measurements that constrain the planet's orbit would be useful for identifying EECs. Indeed, if the planet can be independently detected through RV, and its period constrained with 10% precision, then it can typically be constrained to the habitable zone with $> 80\%$ confidence. However, approximately 25% of EECs remain below our 5 cm/s detection limit, in which case the orbit cannot be established.

Observing the planet multiple times over an orbital period will help to rule out planets outside of the habitable zone with similar confidence, assuming 10% uncertainties on the orbital parameters. Nevertheless, even if a planet can be constrained to the habitable zone, it may yet be too large or small to be habitable. In general, constraints on the orbit and phase will only allow the observer to distinguish EECs from temperate sub-terrestrial planets or ice giants with $\sim 50\%$ confidence.

3.5.3 Do constraints on the mass help to identify EECs?

We find that measurements of the astrometric semi-amplitude θ, when combined with magnitude measurements, can modestly increase the observer's ability to identify EECs. For example, our average confidence for identifying EECs given multiple revisits to determine the orbit (Case 4) improves by about 18% if we include an astrometric measurement or upper limit on the planet's mass (Case 5), and several individual EECs can be identified with very high confidence. These could be the highest priority targets for deeper spectroscopic follow-up.

Under Case 2, if both the measured period and radial velocity semi-amplitude K are used to constrain the planet's properties then EECs can be identified with an average confidence of 52%. If the measured value of K is ignored, however, then this confidence drops to 40%. In other words, the measurement of the planet's mass affords an extra 12% confidence that the planet is an EEC.

Constraining the orbit generally reduces the potential for false positives more than constraining the mass, in part because constraints on the planet's size are available based

Does color information help to identify exo-Earth candidates?
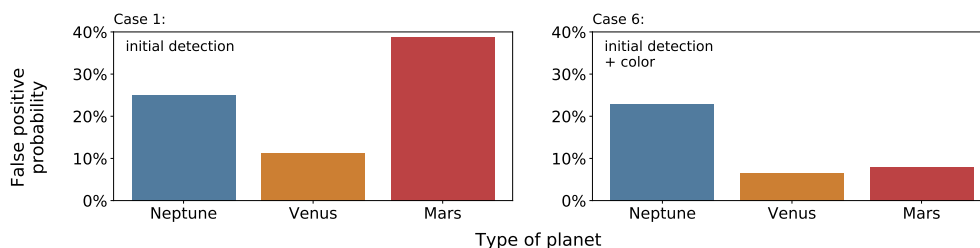
Figure 3.8: We test whether low-to-moderate S/N color measurements would help to identify EECs, assuming that all planets look *approximately* like a solar system analog depending on their class and orbit (Table 3.3 and Figure 3.5). Plotted is the typical inferred probability that the EEC is actually a false positive with a Neptune, Venus, or Mars-type surface and atmosphere. Without color information (left) it is difficult to distinguish Earth analogs from smaller or cooler Mars-like planets, or larger Neptune-like planets. With color information (right), the slope of the optical spectrum provides a useful discriminant between the Earth and Mars, but does little to reduce the ambiguity due to larger Neptunes.

on its brightness alone. On the other hand, the relationship between radius, mass, and composition is more well-understood than the prior distribution of planet albedos (which almost certainly is *not* uniform). An inference about the planet's composition made from its mass could be more reliable than one made from its apparent magnitude.

3.5.4 Do color measurements help to identify EECs?

In Figure 3.8 we plot the average inferred probability that an EEC is similar to the Earth, Neptune, Venus, or Mars for Cases 1 (no color) and 6 (color). We find that color information is useful for distinguishing between spectra with positive versus negative slopes between the UV and visible channels. Specifically, adding a color measurement allows the observer to distinguish between Earth and Martian analogs particularly well - if small planets in the habitable zone tend to look like Mars, color information will be a valuable discriminant. According to Figure 3.6, observing the color *and* constraining the planet's orbit could allow the observer to identify most EECs with $> 80\%$ confidence.

We choose solar system planets as our templates as they cover a relatively broad range of insolations and planet sizes, and with the exception of very hot exoplanets they remain

Figure 3.9: We plot a set of simulated Earth, Neptune, Venus, and Mars analogs in color-color space - each planet's spectrum is modified from the solar system model to simulate physical diversity. We also plot data points from Krissansen-Totton et al. (2016) (Figure 2) who calculate these color ratios for a large diversity of solar system bodies, materials, and exoplanet models. While we consider fewer base models than these authors, our simulated planets cover a similar range in color space.

the only planets for which albedo measurements are presently available. In our set of models shown in Figure 3.5, the Earth stands out due to its modest scattering slope in the optical versus a much stronger feature in Neptune's spectrum, or opposite features for Venus and Mars. Yet a habitable planet does not need to look like the Earth, and indeed there is evidence that during the Archean the Earth had a substantially redder appearance due to organic haze particles (e.g., Arney et al., 2016, 2017). Similarly, the solar system provides no examples of an Earth-sized planet beyond the habitable zone. Here we assume such worlds have a Martian appearance, but this is likely inaccurate for icy worlds or planets with dense atmospheres.

We compare our range of simulated spectra to the models of Krissansen-Totton et al. (2016), who compute optimal photometric bandpasses to distinguish between several different examples of potential exoplanet reflectance spectra. Their optimized bandpasses are 431-531 nm ("blue"), 569-693 nm ("green"), and 770-894 nm ("red") - different than those used in this work. In Figure 3.9 we place our models on a color-color plot similar to Figure 2 in the cited work, along with a subset of the models considered therein. We find that our solar system analogs with simulated physical diversity cover a comparable range in color space, so we believe that we adequately represent a diversity of planet appearances even though our range of base models is limited.

We stress that this result is sensitive both to our prior assumptions and to the bandpasses in which we choose to observe our targets. More work must be done to understand the potential diversity of terrestrial planets and to determine which photometric bandpasses are optimal for distinguishing them from EECs (e.g., Krissansen-Totton et al., 2016). In Section 3.5.9 we discuss how observational constraints on the albedos of potential false positives could be derived within the coming decade. These new discoveries can then be folded into our Bayesian framework, and the results of Case 6 suggest that doing so could allow for the confident distinction between false positives and true EECs on the basis of color information.

3.5.5   Can a debris disk be used to constrain the orbital plane?

We find that measuring the orientation of the debris disk provides relatively little information about the planet's orbit. Specifically, we are on average only $\sim 5\%$ more confident that the observed EECs lie within the habitable zone when we have measured the orientation of the disk. The benefits are slightly greater in cases where the system is observed from a "pole-on" orientation ($|\cos(i)| > 0.5$), in which case the uncertainty in the centroid measurement translates to a smaller uncertainty in the semi-major axis versus the "edge-on" cases ($|\cos(i)| < 0.5$). Nevertheless, unless the centroid precision is much better than 3.8 mas and the observed systems are at least as well-aligned as the planets in the solar system, measuring the disk orientation will generally not allow an observer to constrain the orbit without revisits.

3.5.6   Can EECs be identified given maximal photometric information?

In Case 7 we assume that the observer has revisited the system multiple times to constrain the orbit with three-band photometry and has additionally measured (or placed upper limits on) the astrometric motion of the star due to the planet. This is the most information that could be obtained for the typical system without substantial follow-up or precursor observations, and we find that it would allow the observer to confidently identify most EECs, with an average confidence of 87%. This suggests a promising roadmap toward selecting targets for follow-up, but as in Case 6 this result is dependent on the observer's prior assumptions about exoplanet reflectance spectra.

3.5.7   Would Bayesian prioritization improve follow-up efficiency?

A practical way to frame our results is in terms of follow-up efficiency. A logical next step after detecting a directly imaged planet (and optionally constraining its orbit) would be to search for water absorption in a narrow part of the spectrum to further test its habitability. However, unless a strategy is employed for prioritizing or pruning the target list, a significant amount of time will be spent "following-up" non-habitable planets. Indeed, Kawashima and Rugheimer (2019) have shown that 3-10 hours may be required to confi-

Efficiently surveying planets for spectroscopic H_2O absorption
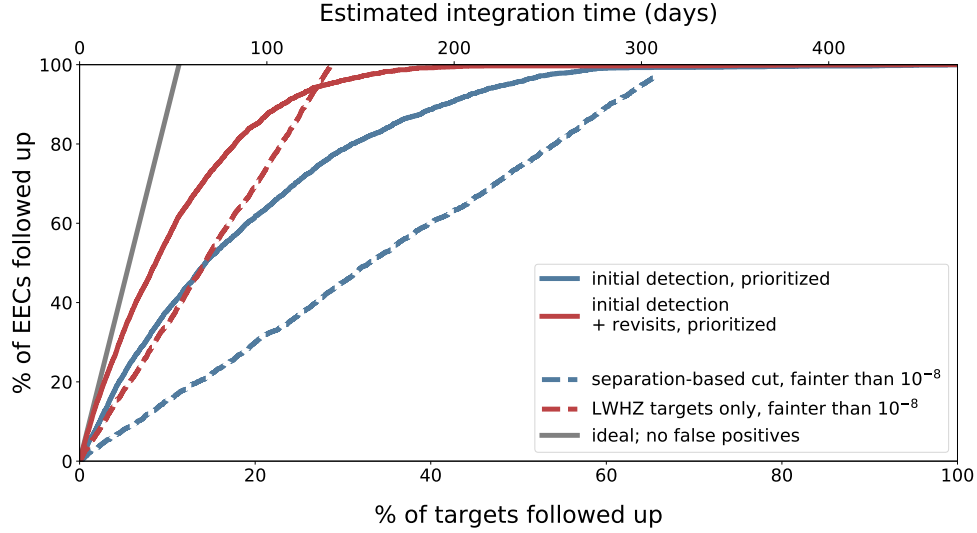
Figure 3.10: The efficiency of different post-detection follow-up strategies (e.g., to search for water absorption), quantified as the percentage of all targets that must be followed up before a given percentage of EECs have been covered; the ordinate is proportional to the amount of time required for follow-up. The gray line represents ideal survey efficiency, where no false positives are re-observed. The colored lines represent strategies that make use of the detection data only (blue) or multiple revisits to establish the orbit (red). The dashed lines are non-prioritized (blind) approaches that first remove very bright or widely-separated planets, or planets whose orbits are constrained to be outside of the LWHZ. The solid lines are prioritized approaches in which we observe the planets that are most likely to be EECs first. The upper axis estimates the amount of time required to search each planet's atmosphere for water absorption, assuming 25 hours of integration time per target.

dently detect H_2O absorption in an exo-Earth atmosphere from 5 parsecs using a 10-meter telescope. For the typical target observed by the 15-meter LUVOIR-A from 15 parsecs, approximately ~ 25 hours may be required.

Our inference framework allows us to prioritize targets based on the likelihood that they are, in fact, exo-Earth candidates. After probabilistically classifying each observed target - including both EECs and non-EECs - we prioritize them by the probability that they are true EECs and submit them for follow-up observations with the most likely candidates first. In Figure 3.10 we summarize the efficiency of this prioritization strategy for

Cases 1 (initial detection only) and 4 (multiple revisits).

We compare our Bayesian approach to two non-prioritized follow-up strategies: first, removing all targets whose projected separations are wider than the maximum greenhouse limit, and second, removing all targets that are found to be outside of the habitable zone after multiple revisits. In both of these cases we also remove targets with a contrast ratio brighter than 10^{-8}. These separation- and magnitude-based cuts exclude $\lesssim 6\%$ of bright or eccentric EECs but several larger or cooler false positives. The remaining targets are then followed-up blindly. Finally, we plot a line representing a perfectly efficient follow-up strategy with no false positives.

We find that prioritizing the targets using our Bayesian framework allows us to re-observe them with a much greater efficiency - using the *same* data - than a blind approach. In fact, a prioritized approach using just the detection data is initially *as efficient* as taking a blind approach after each planet's orbit has been characterized. We can discuss these efficiency gains in terms of integration time. Assuming 450 targets have been discovered in the initial census and 25 hours of integration time are required to search each planet for water, the results in Figure 3.10 suggest that probabilistic target prioritization could reduce the required amount of integration time to follow-up 50% of EECs by 28 days (if the orbits have been precisely constrained) to 95 days (if only the detection data are available). Since this prioritization scheme does not require additional data on the system, it could be naturally folded into the survey strategy of any direct imaging mission.

3.5.8 Is the Bayesian approach always appropriate?

In principle, a Bayesian prioritization scheme should always be superior to a blind follow-up strategy because it leverages additional information about exoplanet demographics. In practice, this prior knowledge is always incomplete and potentially inaccurate, which could make Bayesian prioritization risky in the sense that it might fall prey to unexpected types of false positives. For example, to produce planets for our mock surveys we use the yield estimates of L19, which include more sophisticated treatments of the mission design and a different estimation and extrapolation of *Kepler* occurrence rates - but then we use our own algorithm to probabilistically classify each planet. Therefore, the prior assump-

tions we make when characterizing these simulated planets are not exactly representative of their true parent distribution, which makes our prioritization less efficient.

Indeed, we see in Figure 3.10 that if the planets' orbits are known, then the blind approach to prioritizing targets is *more* efficient than the Bayesian prioritized approach for the last $\sim 5\%$ of EECs. However we argue that this is realistic - the observer's prior assumptions will always differ from reality to a degree - and we note that our approach is still efficient despite this mismatch. For this reason we believe Bayesian prioritization will yield *overall* better results, but observers might prefer a blind approach for low priority targets.

3.5.9 What priors can be improved in the coming decades?

The accuracy of our probabilistic classification depends on the extent and accuracy of existing knowledge of exoplanet statistics. Considering the state of the field ~ 20 years ago, it is reasonable to assume that an observer using this method to interpret observations ~ 20 years from now will base their judgment on better prior assumptions than are currently available. Here we speculate on ways in which the prior inputs to our method could be refined in the coming decades.

3.5.9.1 What is the largest potentially habitable planet?

In our work we assume that planets larger than $\sim 1.6 R_{\oplus}$ will form and maintain volatile envelopes over Gyr timescales, thereby resembling "mini-Neptunes" more than "super-Earths". The most compelling evidence for this comes from density measurements for a limited sample of small planets (Weiss and Marcy, 2014; Rogers, 2015; Chen and Kipping, 2017) and the gap in *Kepler* radius occurrence rates near $\sim 1.7 R_{\oplus}$ (Owen and Wu, 2013; Fulton et al., 2017), but the exact value of this transition radius and its dependence on parameters such as insolation and spectral type requires further research (e.g., Fulton and Petigura, 2018; MacDonald, 2019; Martinez et al., 2019). In the near future, TESS (Ricker et al., 2014) will detect hundreds of small planets orbiting bright stars (Sullivan et al., 2015; Bouma et al., 2017; Barclay et al., 2018; Ballard, 2019). In combination with

precise radii measurements by CHEOPS (Broeg et al., 2013) and ground-based radial velocity measurements, these discoveries will enhance the sample of $1-2R_\oplus$ planets with measured densities. Later on, PLATO (ESA, 2017) will detect several hundred planets smaller than $2R_\oplus$ on orbits as wide as the Earth's, providing statistics for planet radii over a broader range of insolations than *Kepler*. Models of atmospheric evolution can be used to combine these lines of evidence into a more comprehensive understanding of which planets should have non-habitable volatile envelopes (e.g., Owen and Wu, 2013; Lopez and Fortney, 2014; Gupta and Schlichting, 2019).

3.5.9.2 What is the period distribution for planets on wide orbits?

This work relies on extrapolation from *Kepler* occurrence rates for shorter period planets, but ice giants in the habitable zone and beyond could be a significant source of false positives for EEC detection. WFIRST could detect hundreds of wide-orbit planets through its microlensing survey, some with masses lower than the Earth (Barry et al., 2011; Penny et al., 2019). PLATO will also detect a number of transiting planets on orbits as wide as the Earth's (ESA, 2017). Even if these data do not fully probe the relevant range of planet radii and periods, they may provide enough points for interpolation to accurately predict the occurrence of terrestrial planets and ice giants within the habitable zone and beyond.

3.5.9.3 What do the spectra of false positives look like?

Our assumed distribution of planet albedos is the least well-vetted prior assumption in this work, but in Cases 6 and 7 we show that prior knowledge of planet albedos could greatly enhance survey efficiency. The most promising avenue for constraining planet albedos in the next decade is through direct imaging of super-Earths and sub-Neptunes in the habitable zone and on wider orbits with ELTs and WFIRST (e.g., Kasper et al., 2010; Traub et al., 2016; Savransky and Garrett, 2016; Artigau et al., 2018; Weinberger et al., 2018; Akeson et al., 2019). These observations would be valuable for determining the optimum wavelength ranges for discriminating between ice giants and other types of planets.

Several authors have shown that post-runaway atmospheres like Venus' could develop on extrasolar planets comparable in size to the Earth within the inner edge of the habitable zone, which itself depends on the spectral type of the host star and planetary factors such as mass, rotation rate, and ocean coverage (e.g., Kopparapu et al., 2013, 2014; Yang et al., 2014; Kodama et al., 2018). Transit spectroscopy with JWST could reveal the atmospheric composition of a small number of terrestrial planets within the runaway greenhouse limit of low-mass stars (Lustig-Yaeger et al., 2019b). While not directly measuring the albedo, these observations could reveal whether Venus analogs are common, and modeling efforts could reconstruct their likely appearance in reflected light.

3.5.9.4   Testing priors against new data

Finally, it will be possible for observers to validate their priors during the course of the direct imaging survey. As a simple example, if more faint planets are discovered than predicted under the assumed priors, then it is likely that either the occurrence rates underpredict at small radii or there are more low-albedo planets than are present in the solar system. This information could then be forwarded into a probabilistic analysis of which planets are most likely to be Earth-like before substantial time is committed for follow-up observations.

3.6   Conclusions

We have developed a Bayesian framework with which to infer the properties of a directly imaged planet on the basis of limited photometric data, with the primary goal of identifying exo-Earth candidates for deeper spectroscopic follow-up. This framework is dependent on a multitude of priors drawn from observed exoplanet statistics and a few theoretical models. We use it to characterize the ability of future direct imaging missions to identify potentially habitable planets upon their initial detection using only photometry. We determine the key ambiguities involved in this determination and explore potential solutions, such as constraining the orbit through multiple revisits.

Assuming a uniform prior on a monochromatic albedo, we have found that the detec-

tion data alone is not sufficient to determine whether the planet has a potentially habitable size or orbit. In the best cases, a few exo-Earth candidates could be identified with $\sim 50\%$ confidence, but the average EEC would only be identified with $\sim 25\%$ confidence. This translates to a potential false discovery rate of $\sim 75\%$, consistent with previous results (Guimond and Cowan, 2018).

Constraints on the planet's orbit could be achieved through a precursor RV survey or by revisiting the system multiple times. This would allow the observer to constrain the planet to the habitable zone with confidence, but would not eliminate the problem of false positives posed by very small or large planets in the habitable zone. A mass measurement could be somewhat useful for ruling these false positives out, but would by no means be definitive. By revisiting a system multiple times to establish its orbit and measuring the mass astrometrically, an observer could still only distinguish EECs from false positives with a typical confidence of $\sim 65\%$ - but could also identify several individual EECs with high confidence (>90%).

The use of color information to discriminate between EECs and false positives could dramatically reduce these ambiguities, but requires that prior assumptions be made about the possible appearance of planets (e.g., as a function of their size and insolation). While current data and models do not allow such prior assumptions, in principle this could change over the next decade. In our example, we find that by revisiting a planet to establish its orbit and measuring its brightness in three bands, the strong majority of EECs could be identified with confidence.

Even though we are not always able to confidently identify EECs using our method, we show that a target prioritization strategy that leverages a Bayesian approach will be more efficient than a non-probabilistic approach with the same data. Such an approach could reduce the time required for preliminary spectroscopic follow-up by a factor of two.

Our confidence in these results is dependent on our confidence in the priors we have chosen to use, and some of these - such as the radius and semi-major axis distribution of temperate planets around G dwarfs - are based on extrapolation. Many could be improved in the coming decade - such as the relationship between planet radius, mass, and bulk composition through transit and radial velocity observations and advances in planet for-

mation theory, or the reflected spectra of hot and cold sub-Neptunes with JWST, WFIRST, and ELTs. Near-term efforts to improve our prior knowledge will enable future observers to more efficiently survey nearby systems in search of potentially habitable worlds.

3.6.1 Acknowledgements

CHAPTER 4

ACCESS: Ground-based Optical Transmission Spectroscopy of the Hot Jupiter WASP-4b

This chapter has been published as Bixel et al. (2019) and is reproduced here with permission.

Abstract

We present an optical transmission spectrum of the atmosphere of WASP-4b obtained through observations of four transits with *Magellan*/IMACS. Using a Bayesian approach to atmospheric retrieval, we find no evidence for scattering or absorption features in our transit spectrum. Our models include a component to model the transit light source effect (spectral contamination from unocculted spots on the stellar photosphere), which we show can have a marked impact on the observed transmission spectrum for reasonable spot covering fractions ($< 5\%$); this is the first such analysis for WASP-4b. We are also able to fit for the size and temperature contrast of spots observed during the second and third transits, finding evidence for both small, cool and large, warm spot-like features on the photosphere. Finally, we compare our results to those published by Huitson et al. (2017) using *Gemini*/GMOS and May et al. (2018) using IMACS, and find that our data are in agreement.

4.1 Introduction

WASP-4b is a $1.4 R_J$ hot Jupiter orbiting a G7V star with a period of 1.34 d, equilibrium temperature of ~ 1700 K, and transit depth $(R_p/R_s)^2 = 2.4\%$ (Wilson et al., 2008). It has been observed in transit over three dozen times, offering strong constraints on its orbit (Hoyer et al., 2013), the spot activity and relative rotation of its host star (Sanchis-Ojeda et al., 2011), and also placing upper limits on its transit timing variation amplitude (Nikolov et al., 2012).

Beerer et al. (2011) used *Spitzer* to observe the planet's secondary eclipse and place constraints on the temperature profile of its atmosphere, and they conclude that the evidence is consistent with either a weak temperature inversion or none at all. However, even stronger detections of thermal inversions using *Spitzer* have later been called into question (e.g., HD 209458b, Diamond-Lowe et al., 2014). The evidence for an inversion therefore remains marginal, leaving us with little insight about the composition of the planet's upper atmosphere.

Given the tight constraints on its transit depth and orbital properties, WASP-4b is a natural target for spectroscopic transit observations. Transit spectroscopy can be used to constrain the composition and structure of the planet's upper atmosphere and test for the presence of clouds, scattering hazes, and atomic or molecular absorbers (e.g., Seager and Sasselov, 2000; Brown, 2001; Hubbard et al., 2001). These observations also offer insight into the stellar photosphere, i.e., through the measurement of star spot temperatures within the transit chord (e.g., Pont et al., 2008; Sing et al., 2011; Béky et al., 2014). However, signals from the photosphere can be degenerate with those from the planet's atmosphere, leading to contrasting interpretations of planetary and stellar origins for features in optical transmission spectra (e.g., Pont et al., 2013; McCullough et al., 2014). It is therefore critical to demonstrate methods for accounting for this degeneracy as the field moves toward smaller targets and increasingly precise observations (e.g., Rackham et al., 2018).

Transit spectroscopy of WASP-4b has been attempted with *HST*/WFC3 (Ranjan et al., 2014), but was unsuccessful due to detector saturation. A *Gemini*/GMOS optical transmission spectrum of WASP-4b has been published by Huitson et al. (2017), who mea-

sure a nearly uniform opacity from 440-940 nm, suggesting the presence of high-altitude clouds and a possible sodium absorption feature. More recently, May et al. (2018) have published a transmission spectrum using *Magellan*/IMACS, and have similarly found no evidence for spectral features.

As part of the Arizona-CfA-Católica-Carnegie Exoplanet Spectroscopy Survey (AC-CESS, Rackham et al., 2017), we have observed four transits of WASP-4b with *Magellan*/IMACS. We have previously demonstrated the use of this instrument for transit spectroscopy using our custom data reduction pipeline (Jordán et al., 2013; Rackham et al., 2017; Espinoza et al., 2019). In this paper, we present an optical transmission spectrum from 450-900 nm. We interpret our results using a Bayesian retrieval code introduced in Espinoza et al. (2019) and find no evidence for scattering or absorption features. We also fit for the size and temperature of photosphere features occulted during the second and third transits, and derive corrections for their effects on the transit spectrum. Finally, we compare our findings with those of Huitson et al. (2017) and May et al. (2018).

4.2   Observations

We conducted spectroscopic observations of WASP-4 on the nights of 24 September 2013, 17 October 2013, 14 August 2015, and 26 September 2015 (hereafter Transits 1–4) using the Inamori-Magellan Areal Camera & Spectrograph (IMACS, Dressler et al., 2011) on the 6.5m Magellan-Baade telescope at Las Campanas Observatory in Chile. We observed using multi-slit masks in the f/2 mode with 2x2 binning (0.4"/px). Observations of HeN-eAr and quartz calibration lamps before and after the observations allowed for wavelength calibration and flat-field correction. The key parameters of our observations are listed in Table 4.1.

4.2.1   24 September 2013, 17 October 2013, and 14 August 2015

On the first three nights we used a setup consisting of a 400–1000 nm spectroscopic filter, a 300 lines/mm grism with a blaze angle of 17.5 degrees, and 10" wide by 20" long spectral slits for the target and reference stars. Twelve reference stars were observed,

although only one was used in the final data reduction for reasons discussed in Section 4.3.3. Most of the stellar spectra were dispersed across two chips. While the lunar sky background was minimal on the first and third nights, it was substantial during the second, and we take this into account in our data reduction pipeline. Finally, the observations on 24 September 2013 (Transit 1) did not commence until shortly after ingress, so we were only able to observe a partial transit.

4.2.2   26 September 2015

Our instrument setup for the fourth observation was modified from the previous three. We used a 570–980 nm order-blocking filter for the purpose of eliminating higher-order interference towards red wavelengths, 10" wide by 10" long slits, and a 150 lines/mm grism with a blaze angle of 18.8 degrees. This setup allowed for more tightly dispersed spectra that fall on a single chip, thereby reducing detector-to-detector variations in the spectra and avoiding chip gaps. The moon was in full phase and separated from the target by 40 degrees, contributing significantly to the ambient sky background, which we again account for in our data reduction pipeline.

Table 4.1.   Instrument Setup and Transit Model Characteristics for Each Observation

Transit	Date (start of night)	Filter	Grism (l/mm)	Airmass	R_p/R_s (white light)	Scatter / Photon noise
1	24 September 2013	Spectroscopic f/2	300	< 1.3	$0.1528^{+0.0012}_{-0.0011}$	4.9
2	17 October 2013	. . .	. . .	< 1.1	$0.1537^{+0.0008}_{-0.0007}$	7.8
3	14 August 2015	. . .	. . .	< 1.7	$0.1544^{+0.0002}_{-0.0002}$	2.9
4	26 September 2015	WBP 5694-9819	150	< 1.2	$0.1565^{+0.0004}_{-0.0005}$	6.2

Note. — All four observations made use of the IMACS f/2 camera with multi-slit masks. The ratio in column 7 factors in the brightness of the target and single reference star, but not the sky background.

4.3 Data reduction

We reduce the raw data using our custom Python-based pipeline, which has been used for similar observations of WASP-6b (Jordán et al., 2013), GJ 1214b (Rackham et al., 2017), and WASP-19b (Espinoza et al., 2019). In the following paragraphs, we give a brief overview of the pipeline functions. A more detailed review can be found in Jordán et al. (2013) and Rackham et al. (2017); the only more recent addition to the pipeline is the correction for the non-uniform sky background described in Section 4.3.1.

We use quartz lamp images taken with the same configuration as the science images to apply a flat-field correction, and we calculate the bias offset from the overscan region of each chip. We use full-frame flats taken without a mask or grism to identify bad pixels, and we identify cosmic rays on the stellar spectra using a 3σ threshold at each row along the direction of spectral dispersion. To trace the stellar spectra, we calculate the centroid in each row along the dispersion direction and fit a second-order polynomial to the centroid values. For Transits 1 and 2, we assume the background is uniform in the spatial direction, measure it using the outermost 14 px in each row along the dispersion direction, and subtract. For Transits 3 and 4, the background is not spatially uniform, so we subtract it using the method outlined in Section 4.3.1. Finally, to extract the spectra we use the optimal extraction algorithm outlined by Marsh (1989), which involves fitting a third-order polynomial to the spectral profile at each row along the dispersion direction, then using that profile to weight each pixel when summing the flux.

We use HeNeAr arc lamp exposures taken before and after the transit observations to calculate the wavelength solutions for each star, which convert the pixel coordinates along the dispersion direction to wavelength values. We manually identify prominent emission lines and use a sixth-order polynomial to fit the wavelength solution; given the large number of lines used, the danger of over-fitting is minimal. The marked lines are iteratively rejected based on their residuals to the fit until the residuals are below ~ 0.05 Å.

Our spectra drift in the dispersion direction over the course of the night, resulting in frame-to-frame offsets in the wavelength solution. To solve this, we cross-correlate

the first spectrum of a star with each of the subsequent spectra to determine the shift in wavelength, then fit a third-order polynomial to this shift as a function of time and use it to correct the wavelength solution in each frame.

The output of the pipeline is a set of reduced wavelength-binned and integrated ("white") light curves for the target and reference stars.

4.3.1 Non-uniform sky background

In some of our IMACS data sets for this and other ACCESS targets, we have noticed that spatially uniform sources of light inherit a non-uniform profile in the spatial direction once the mask and dispersive element are in place. This effect also widens the spectra of point sources. We have attributed this to internal scattering within the instrument, and have found that it is more common in newer images: of the four data sets for WASP-4, only Transits 3 and 4 are affected. We measure the extent of the effect using our flat field images, finding that the profile of spatially uniform light peaks at the slit center and appears $\sim 10\%$ fainter at the slit edges. Unless we account for this, we will underestimate the sky background when we extract the stellar spectra.

For the two affected data sets, we use quartz lamp exposures taken at the beginning of the night to model the scattered light profile. The peaked profiles are not consistent with common symmetrical functions (e.g., Gaussian), nor are they well-described by a classical high-order polynomial, because high-order polynomials commonly fail at the edges of their fitted intervals (an effect known as Runge's phenomenon). Instead, we use Chebyshev polynomials, which are not as susceptible to this effect, and find that sixth-order polynomials are sufficient to match the data.

The polynomials are fitted independently for every row along the dispersion direction of every slit in the flat field images. Then, in each science image we fit the amplitudes of the polynomials using the background level in the outermost 14 px in each row along the dispersion direction (8 px for Transit 4 due to shorter slits). Finally, we subtract the fitted polynomial from each row.

4.3.2 Flat field correction

Applying the flat-field correction for Transits 1, 3, or 4 does not significantly change our binned light curves, or the transit spectra that we derive in Section 4.4. However, applying the correction for Transit 2 introduces $\sim 2\%$-level, non-linear time-dependent trends into out-of-transit baseline flux of the binned light curves blueward of 530 nm. These trends remain even when normalizing by the reference star as discussed in the following. Therefore, for consistency, we choose not to apply the flat-field correction to any of our data sets.

4.3.3 Reference star selection

The shape of the target's light curve is complicated by instrumental and atmospheric effects, such as changes in airmass or transparency. To calibrate out these effects, we simultaneously observed 12 reference stars of comparable optical apparent magnitudes and color ratios using multi-slit masks. Of these, two spectra reached the saturation limit of the detector and were not usable.

We use our highest quality data set (Transit 3) to determine which of the ten remaining reference stars to use in our light curve analysis. Our primary consideration is the shape of the out-of-transit baseline for the target light curve when normalized by each star. Eight of the stars leave residual long-term trends at the $\sim 1\%$ level. Two stars leave trends at the $\sim 0.1\%$ level, and they happen to be the only two stars occupying the same pair of detector chips as the target. The point-to-point scatters of the baseline points for the two resulting light curves are 0.4 and 0.6 mmag, which are both smaller than the scatters for the eight light curves with larger long-term trends; we therefore further restrict our set of potential reference stars to these two. We perform a similar analysis of the two remaining stars using our three other data sets, and discover that one of the two stars reliably produces a flat baseline, while the other star introduces trends at the $\sim 0.5 - 1\%$ level. Therefore, we discard this star as well.

We are left with a single reference star, 2MASS J23341836-4204509, that reliably produces a flat out-of-transit baseline with a low point-to-point scatter in all of our data

sets. We note that this reference star was also used in a similar analysis by Huitson et al. (2017). The spectral type of the reference star has not been previously published, but using parallax measurements from Gaia DR2 (Gaia Collaboration, 2018) we calculate that it is intrinsically brighter than the target by $\sim 80\%$ from 400 - 1000 nm. While this wavelength range does not capture all of the emitted light, it does cover the emission peak for late F and all G stars. Using a simple $L_* \propto M_*^{3.5}$ scaling relation, and assuming that our wavelength range covers most of the emitted light, this suggests that the reference star is $\sim 20\%$ more massive than the G7V, 0.85 $M_\odot$ target (Gillon et al., 2009), consistent with an early G or late F dwarf.

4.3.4 Effects of atmospheric dispersion

Magellan-Baade is equipped with an atmospheric dispersion compensator (ADC), which corrects for the effects of differential atmospheric refraction as the airmass of our target changes over the course of the night. If left uncorrected, differential refraction would affect our stellar wavelength solutions in a non-linear manner, "stretching" or "shrinking" the spectra over the course of the night. We test the accuracy of the ADC correction by measuring the distance in calibrated wavelength space between Na and Hα in the target and reference spectra as a function of time, which we calculate to vary by no more than 1 Å over the course of the night. Furthermore, the difference between the measured distance for the target and reference stellar spectra changes by no more than 0.5 Å, and less than 0.1 Å on nights with higher quality data. Since the long-term change in this measurement is much lower than the frame-to-frame scatter (up to 5 Å), we reason that any effect on the measured transit depth will be negligible.

4.3.5 Observing efficiency

During Transits 1 and 2, we opted for longer, low-noise readout modes and short exposure times to avoid detector non-linearity. As a result, we spent only $\sim 20\%$ of our time gathering light. By comparison, we achieved an efficiency of $\sim 70\%$ and $\sim 50\%$ during Transits 3 and 4, respectively, as we chose the fastest readout mode and took longer

exposures. The difference in read noise between the fastest (5.6 e-) and slowest (2.8 e-) modes is negligible compared to the Poisson noise ($>$100 e-), and as we show in Section 4.3.8, the detector is sufficiently linear for our purposes provided the pixel values remain near or below half-well.

Based on this experience, we advise that observers prioritize longer exposure times with fast readouts for transit spectroscopy with IMACS.

4.3.6 Sky background

Transits 2 and 4 suffer from high sky background values due to the target's proximity to the full-phase moon. The effects can be measured by comparing the scatter of the residuals in Figure 4.2. Scaled for exposure time, the magnitudes of the scatter for Transits 1 and 3 agree to within 15%, while the scatter of Transit 2 is $\sim 50\%$ larger. The scatter of Transit 4 is similarly $\sim 50\%$ larger, although Transit 4 was observed in a redder filter with less sky contamination.

4.3.7 Slit losses

We consider whether slit losses may have been significant during our observations. For Transits 1-3, the slits were 10" wide and 20" long, while for Transit 4 the slits were 10" wide and 10" long. We perform a least-squares fit of a Moffat profile to the point spread function of a bright but unsaturated star in each of the field acquisition images, then integrate to determine how much light would lie outside of the slit if placed over the star. For every night, the fraction of light lost is more than an order of magnitude lower than the Poisson noise.

4.3.8 Detector linearity

We observed Transits 3 and 4 with longer integration times to reduce the relative readout overhead and improve our signal-to-noise ratios. However, this increased the maximum pixel values of the target spectrum from $\sim 15,000$ to $\sim 30,000$ ADU, nearly half of the full well value ($\sim 65,000$ ADU). This raises the question of whether the detector's

response could have become non-linear (i.e. the gain was not uniform over the range of pixel values), which could bias the mean-subtracted transit spectra which we extract later in Section 4.4.

In principle, non-linearity should not be an issue for measuring the relatively small changes in transit depth from bin to bin, as the gain should not be expected to change over the small range of changing values. However, the pixel values at the peak of the target spectrum varied by up 15% from frame to frame and 40% over the course of each night, due mostly to seeing variations that changed the point spread function. Therefore it is worthwhile to determine the threshold beyond which non-linearity could significantly affect our results.

To do so, we modify the raw data for Transit 3 to reflect a 0.1% or 1% linear increase in the gain over the range of 0-35,000 ADU. We then reduce the modified data and extract the mean-subtracted transit spectra for both gain prescriptions, which we plot alongside the original spectrum in Figure 4.1. We find an average effect of 0.07σ per bin for a 0.1% increase in gain, and 0.23σ per bin for a 1% increase.

So long as the detector is linear to 0.1% from zero to half-well, non-linear effects should have a negligible impact on our results. Even in the case of 1% non-linearity, the effect is $\ll 1\sigma$ except in a few bins. We are therefore confident that our results are robust to realistic levels of non-linearity, and we advise that future observers target a similar range of peak pixel values ($\sim 35,000$ ADU) so as to optimally balance observational efficiency and detector linearity.

4.4   Light curve modeling

We model our integrated and spectroscopic light curves using the analytic models introduced by Mandel and Agol (2002), and marginalize over the parameter space with a Markov-Chain Monte Carlo (MCMC) algorithm previously detailed in Rackham et al. (2017). As in previous works, we fit for the limb-darkening coefficients in order to account for any biases that might arise due to our imperfect knowledge of the intensity profile of the photosphere (Espinoza and Jordán, 2015). As WASP-4b has been studied
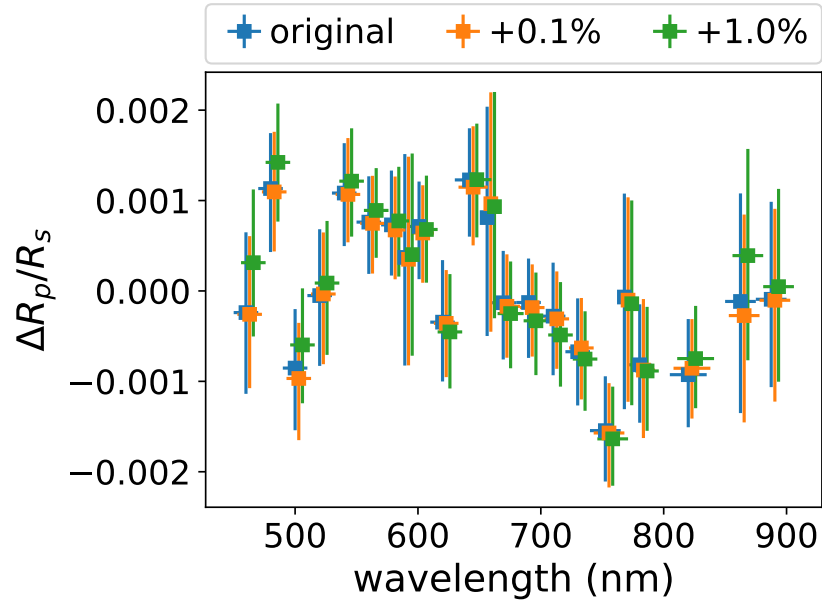
Figure 4.1: The mean-subtracted transit spectrum of Transit 3 with different gain modifications. We calculate and apply a gain enhancement to each pixel in the raw data, assuming the gain increases linearly by 0%, 0.1%, or 1% from 0-35,000 ADU. Provided the detector is linear to $< 1\%$ over this range, the mean-subtracted transit spectrum is largely unaffected.

extensively through photometric observations, we hold the orbital parameters fixed to the mean values in Table 4.2.

The fitted parameters for the transit model include the planet-to-star radius ratio (R_p/R_s), two parameters for a quadratic limb-darkening law, which are sampled according to the method detailed in Kipping (2013a), and the mid-transit time. The baseline flux is modeled as a second-order polynomial; this decision is discussed in more detail in Section 4.4.4.

We use the likelihood function of Carter and Winn (2009), Equation 41, in which the noise in the light curve is parameterized by two free noise parameters, σ_w and σ_r. σ_w describes the amplitude of uncorrelated ("white") sources of noise (e.g. photon noise), while σ_r describes the amplitude of the correlated ("red") noise, which is modeled as a superposition of time-localized oscillating signals known as wavelets. The power spectral density of the red noise is modeled as $S(f) \propto 1/f^\gamma$; following the authors' example, we set $\gamma = 1$. We calculate the wavelet functions and likelihoods using our own Python module[1].

We first fit a model to the white light curve of each transit to determine the mid-transit time, then fit each binned light curve independently with the mid-transit time fixed. The fitted white light curves for each night are shown in Figure 4.2, and the fitted binned light curves are presented in Appendix Figures A.1 through A.4.

Note that two of the transits exhibit spot-crossing features that have been excluded from the fit. To determine which data points to exclude, we first conservatively remove data near the spot and fit a model to the white light curve, then fit a one-dimensional Gaussian model to the shape of the spot in the residuals, and finally exclude only those data points that lie within two standard deviations of the fitted mean. For a more detailed analysis of these spot features, see Section 4.5.

The primary output of our fitting routine is a transmission spectrum for 19 independently fitted continuum bins, each approximately 20 nm in width, and three narrower bins centered on possible absorption features, including Na D, Hα, and the K I doublet (767/770 nm). We select the bin width to be as small as possible while maintaining signal-to-noise ratios of at least a few hundred in each bin and data set. We attempt to

---

[1]https://github.com/nespinoza/flicker-noise

keep the bin sizes consistent across the spectrum, but some bins have been adjusted to accommodate the three narrow bins and two chip gaps. The spectra for each transit are presented in Figure 4.3, covering wavelengths from 450–900 nm in the first three nights, and 570–900 nm for the fourth night. We have no data from 790-805 nm and 835-850 nm as these correspond to the location of the detector chip gaps in the target and reference star spectra.

4.4.1 Combining data from separate nights

The average values of R_p/R_s in the spectrum of Transits 1-3 agree to within 1.5σ, but Transit 4 disagrees at the $\sim 5\sigma$ level. This mirrors the deviation in the fitted orbital parameters for Transit 4 discussed in Section 4.4.3, and the causes are likely the same.

Regardless of the reason for this disagreement, we are primarily interested in the *change* in R_p/R_s with wavelength. We subtract the weighted mean $\overline{R_p/R_s}$ from the individual spectra for Transits 2-4, then calculate the average values of $\Delta R_p/R_s$ in each bin, weighted by the uncertainties for each night. The combined transmission spectrum is presented in Figure 4.4.

Due to the large uncertainties and incomplete baseline coverage for Transit 1, we do not include its spectrum in the combined result; see Section 4.4.6. We also correct the spectrum of Transit 3 for known stellar contamination before incorporating it into the combined spectrum, as detailed in Section 4.4.7.

4.4.2 Excluded bins

Two of the fitted bins have been excluded from our primary figures and results: the 450-470 nm bin of Transit 2, and the 875-900 nm bin of Transit 4. The values for both bins deviate by $> 3\sigma$ from the average value of their respective spectra, and both bins are at the low S/N ends of the stellar spectrum. For reference, these bins are included in Appendix Figures A.2 and A.4 and Table A.1.

The 450-470 nm bin of Transit 2 lies $\sim 3\sigma$ *below* the average value of the spectrum; we can think of no astrophysical explanation for this effect. Furthermore, the likelihood

Table 4.2. Relevant Previously Measured Properties of WASP-4 and Its Companion,
With 1σ Uncertainties

Parameter	Value	Reference
WASP-4		
R_s ($R_\odot$)	$0.873^{+0.036}_{-0.027}$	Gillon et al. (2009)
M_s ($M_\odot$)	$0.85^{+0.11}_{-0.07}$	. . .
[Fe/H]	$-0.03^{+0.09}_{-0.09}$	. . .
log(g) (cgs)	$4.487^{+0.019}_{-0.015}$	. . .
T_s (K)	5540^{+55}_{-55}	Maxted et al. (2011)
WASP-4b		
P (d)	1.33823204	Hoyer et al. (2013)
R_p/R_s	$0.15445^{+0.00025}_{-0.00025}$	. . .
i (deg)	$88.52^{+0.39}_{-0.26}$	. . .
a/R_s	$5.463^{+0.025}_{-0.020}$	. . .
e	≈ 0	Beerer et al. (2011)
R_p (R_J)	$1.395^{+0.022}_{-0.022}$	Hoyer et al. (2013)
M_p (M_J)	$1.237^{+0.021}_{-0.021}$	Winn et al. (2009)

of a statistical 3σ outlier across all of our 82 bins is $\sim 0.3\%$. Most likely, our polynomial model is inadequate to describe the low S/N systematic trends in this bin.

The 875-900 nm bin of Transit 4 lies $\sim 4\sigma$ above the weighted mean. While this wavelength range does correspond with water absorption, such an offset is not observed in the other spectra. Furthermore, the value of this bin depends strongly on whether a linear or quadratic polynomial is used to model the statistics, as discussed in Section 4.4.4. Again, it seems most likely that our systematics model is inadequate for this low S/N bin.

4.4.3 Fitting orbital parameters

We also use our MCMC code to simultaneously fit for the transit depth, the inclination (i) and the semi-major axis (a/R_s), while keeping the eccentricity and period fixed. The results of these fits are listed in Table 4.3. The parameters from Transits 1-3 are generally consistent with those measured by Hoyer et al. (2013), although the values from the partial Transit 1 are poorly constrained.

The parameters from Transit 4 deviate $2-3\sigma$ from the previously measured values. It is not clear why this is the case. Stellar variability may be a partial contributor; as discussed in Section 4.6.2, WASP-4 is variable at the 6 mmag level, and this can lead to changes in the apparent transit depth. The deviation may also be related to losses due to the shorter slits employed during Transit 4, as discussed in Section 4.3.7.

Due to this this large deviation, we also try fitting the spectrum of Transit 4 by fixing the inclination and semi-major axis to the fitted values, rather than those in the literature. The result is only a $\ll 1\sigma$ change in the binned values for $\Delta R_p/R_s$, so we believe the spectrum for this night is robust despite the disagreement in the orbital parameters. For consistency with the other data sets, we keep the parameters fixed to the literature values in the remaining sections.

4.4.4 Systematics model

We find that the trends that remain in the target light curves after dividing by the reference star light curves correlate with airmass, which tends to reach its minimum near the center of the transit. Nevertheless, no physically motivated (e.g., exponential) function of airmass is able to describe this trend better than a simple second-order polynomial in time, so we opt to use time-dependent polynomials to characterize our long-term systematics.

The long-term trends in many of our binned light curves cannot be adequately described by a linear polynomial in time, so we use a quadratic polynomial instead. In principle, any degeneracy between the transit depth and the polynomial value should be incorporated into our MCMC uncertainties. To measure the impact of the model choice on our results, we apply both linear and quadratic systematics models and compare the
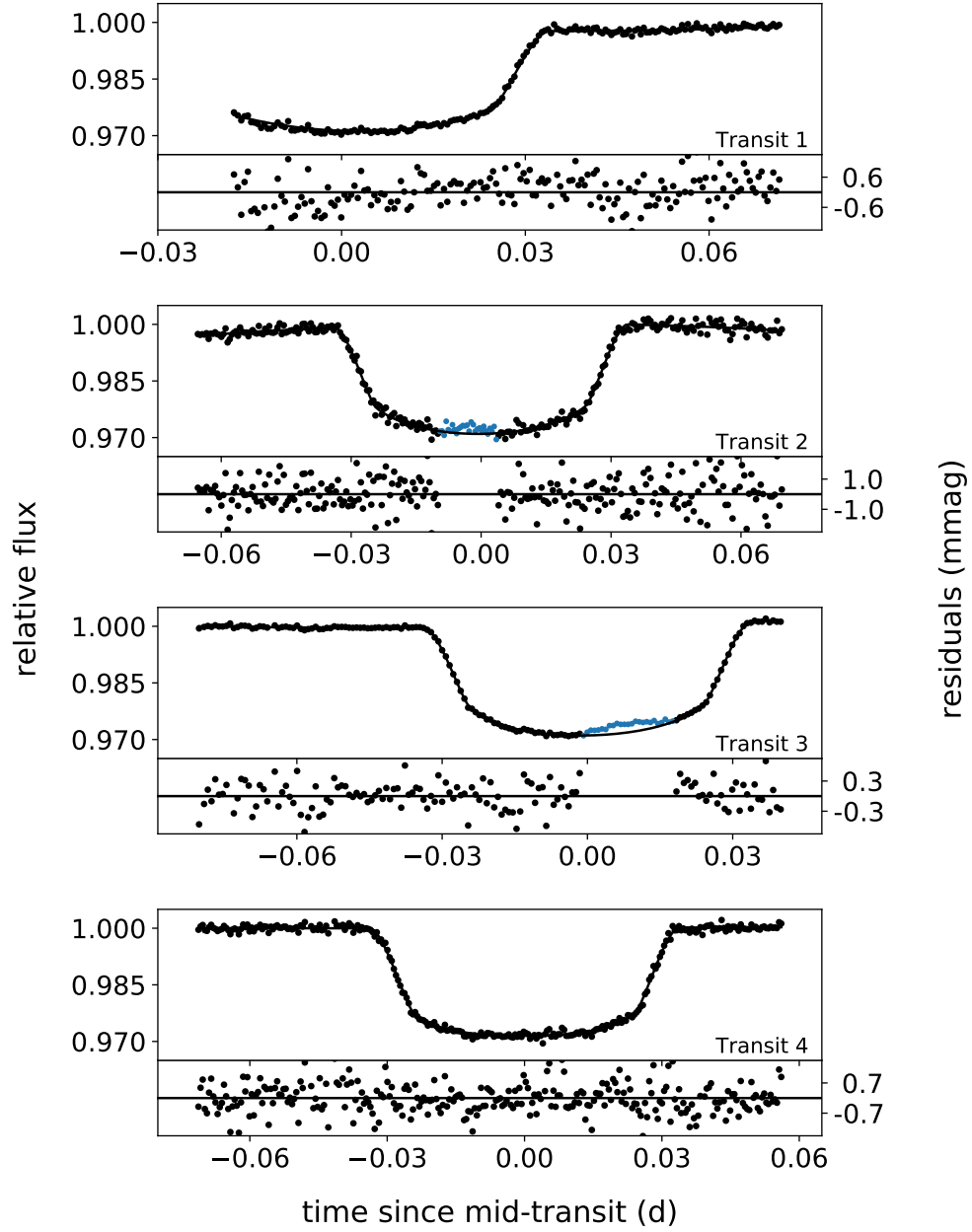
Figure 4.2: White light curves with transit models and residuals (1σ values marked). Two of the light curves featured spot-crossing events (blue) that are not included in the fit. The long-term trend is modeled by a second-order polynomial.
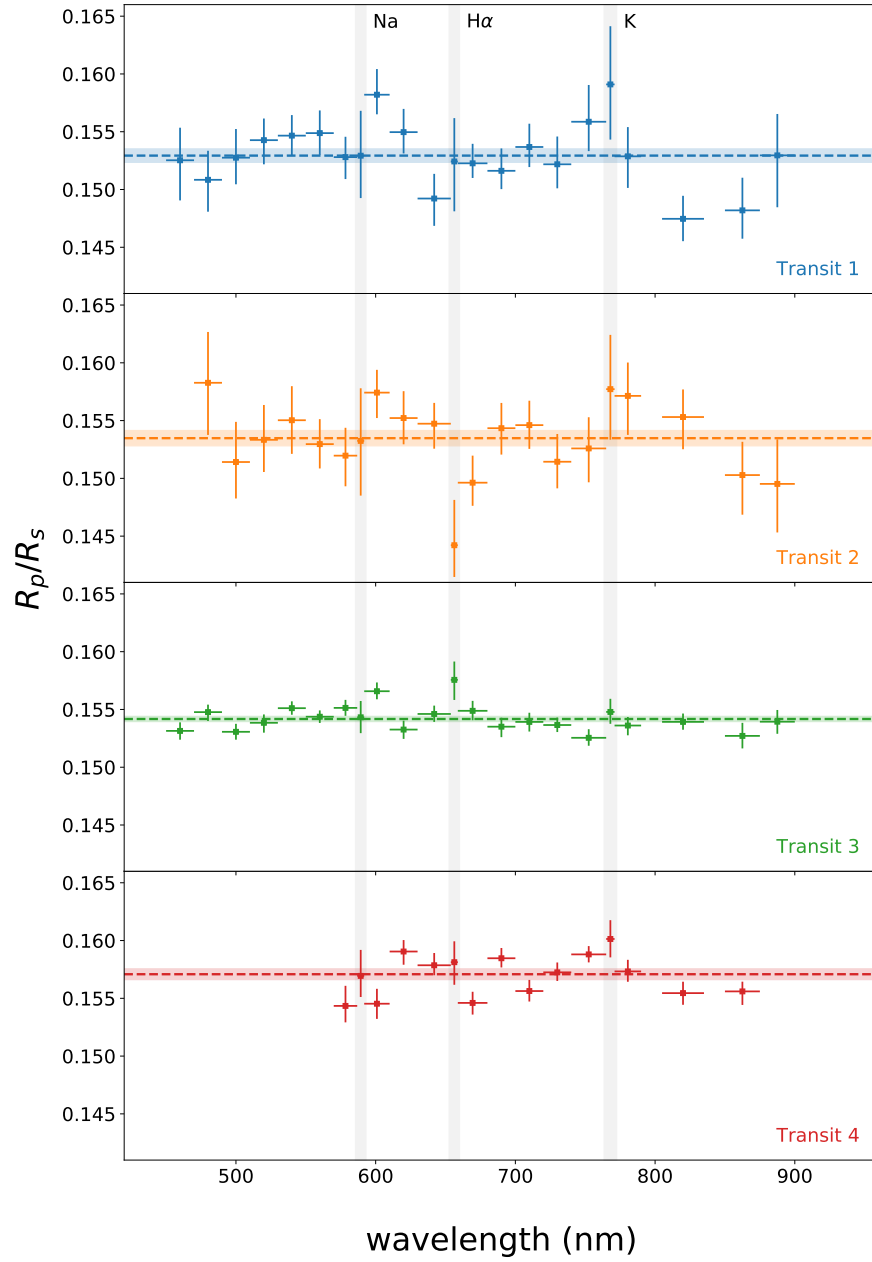
Figure 4.3: Transmission spectra from each night, with no corrections or offsets applied. The dashed lines and shaded regions represent weighted average values with $\pm 1\sigma$ uncertainties. The wavelengths of potential atomic features are highlighted.

Table 4.3.   Fitted Orbital Parameters When Modeling the Planet Radius, Inclination, and Semimajor Axis Jointly

	R_p/R_s	a/R_s	i (deg)
Hoyer+13	$0.15445^{+0.00025}_{-0.00025}$	$5.463^{+0.025}_{-0.020}$	$88.52^{+0.39}_{-0.26}$
Transit 1	$0.15386^{+0.00135}_{-0.00174}$	$5.386^{+0.088}_{-0.157}$	$89.33^{+0.47}_{-0.73}$
Transit 2	$0.15617^{+0.00164}_{-0.00167}$	$5.439^{+0.070}_{-0.092}$	$87.90^{+1.24}_{-0.85}$
Transit 3	$0.15388^{+0.00045}_{-0.00047}$	$5.473^{+0.021}_{-0.038}$	$88.94^{+0.53}_{-0.52}$
Transit 4	$0.15654^{+0.00067}_{-0.00065}$	$5.520^{+0.026}_{-0.030}$	$89.46^{+0.37}_{-0.60}$

Note. — For comparison we include the values derived by Hoyer et al. (2013).

values of $\Delta R_p/R_s$ for each.

For Transits 1-3, the average absolute change in $\Delta R_p/R_s$ between the linear and quadratic models is approximately 0.5σ per bin, and no larger than 1.6σ in any bin. For Transit 4 the average change is approximately 1.5σ per bin, but the reddest bin (which has the lowest S/N) changes at the 3σ level. Due to this large inconsistency between systematics models, we exclude the reddest bin from our results.

Some authors use the residuals to their white light curve models to calibrate out common (i.e., wavelength-independent) systematics from each bin (e.g., Sing et al., 2013; Nikolov et al., 2015; Huitson et al., 2017). In our case we find that such a correction is unwarranted. By eye, it is not apparent that our binned light curves share any common systematics (see Figures A.1-A.4 in the Appendix). Furthermore, when we apply the common mode correction, we find that the impact on the resulting transit spectra is negligible compared to the uncertainties.
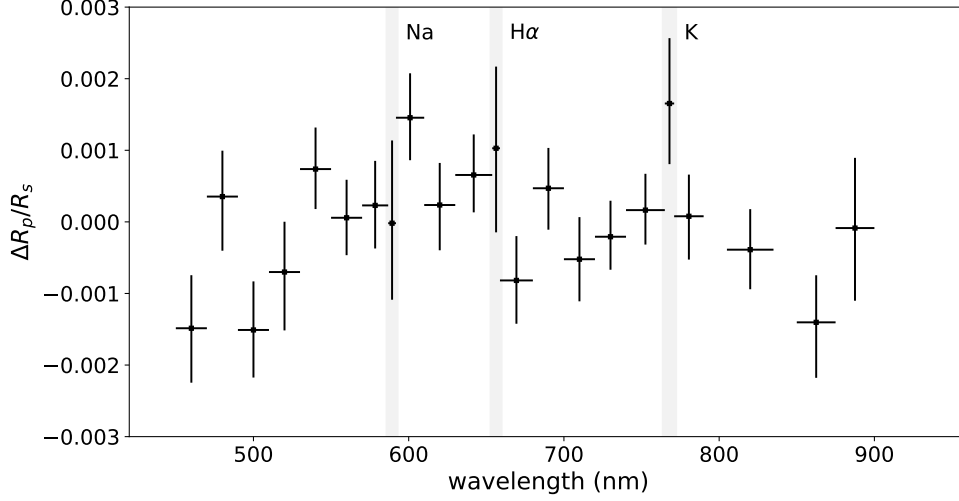
Figure 4.4: Combined transmission spectrum from Transits 2-4, with the weighted average value subtracted from each night. Narrow bins centered on potential atomic features are highlighted, and the remaining (continuum) bins are black. The spectrum of Transit 3 is corrected for the effect of an occulted spot as described in Section 4.4.7 before it is incorporated into the combined result. Transit 1 is excluded due to poor phase coverage.

4.4.5 Red noise

As discussed in Section 4.4, we fit a parameter σ_r to characterize the level of correlated ("red") noise in our data. While the ratios in Table 4.1 suggest a large amount of systematic noise, this does not necessarily imply *correlated* noise; the systematic noise may still be captured by the white noise parameter.

To assess the level of red noise in our data, we construct Lomb-Scargle periodograms of the residuals to each binned light curve model, then use bootstrap resampling to calculate the false alarm probability (FAP; e.g., VanderPlas, 2018) for signals with periods between the exposure times and the duration of our observations. Figure 4.5 shows an example periodogram for one of the bins of Transit 3. We find that only 6 out of 82 binned light curves possess signals with FAP < 10%, and only 2 have FAP < 1%. This suggests that there are few strong periodic signals in our light curves. Nevertheless, we
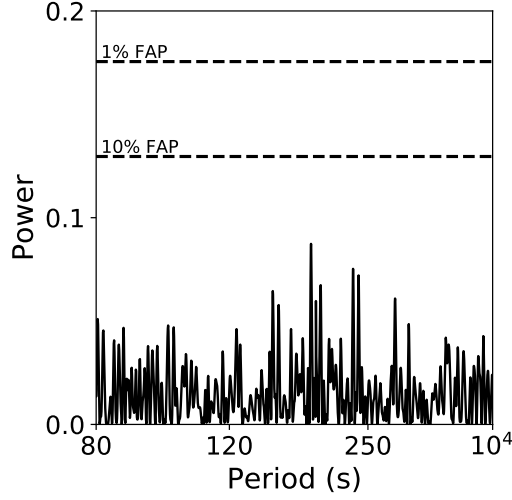
Figure 4.5: Lomb-Scargle periodogram of the residuals to the light curve model for the 610-630 nm bin of Transit 3. The lack of strong signals demonstrates that this light curve does not have strong periodic correlated noise across this range of frequencies.

conservatively choose to apply the red noise parameterization, which results in a $\sim 15\%$ increase in our uncertainties.

4.4.6 Effects of incomplete phase coverage

The light curves of Transits 1-3 all suffer from a partial lack of phase coverage, due either to poor weather, observing window constraints, or spot-crossing events that must be masked in the model fit. In principle, our MCMC approach to fitting the model should adequately incorporate this lack of information into the uncertainties on R_p/R_s. However, it is worthwhile to estimate the extent to which this missing information inflates our uncertainties.

The full light curve of Transit 4 allows us to investigate the effects of poor phase coverage in the previous three transits. Specifically, we remove data points from the binned light curves of Transit 4 to mimic the phase coverage of Transits 1-3, then fit for the transit depth and calculate the resulting increase in the uncertainties on the binned values of R_p/R_s. Our results are summarized as follows:

- Mock Transit 1: We remove the pre-ingress baseline and most of the first half of the transit. This results in a 150% increase in uncertainties, as well as a 4.5σ decrease in the wavelength-averaged radius. The $\Delta R_p/R_s$ values change at the $> 1\sigma$ level, and the change appears to be wavelength dependent, introducing a slope on the order of $\sim 1\sigma$ from blue to red.

- Mock Transit 2: We remove the points near mid-transit that correspond to a spot-crossing event. This results in a 50% increase in uncertainties, and a 0.5σ decrease in the average radius. The $\Delta R_p/R_s$ values change at the $\lesssim 0.2\sigma$ level.

- Mock Transit 3: First, we remove most of the post-egress baseline, which results in a 50% increase in uncertainties with no considerable change in the average radius, and $\lesssim 0.5\sigma$ changes in $\Delta R_p/R_s$. Next, we exclude only those points that correspond to the spot crossing event in Transit 3, finding effects of similar magnitude. Removing both portions of the light curve, however, does not increase the uncertainties any further.

The results for Transit 1 lead to concerns over whether the lack of a pre-ingress baseline could introduce a bias into the shape of our transit spectrum. Furthermore, the large uncertainties mean that this transit contributes little to the combined spectrum. We therefore opt to exclude the spectrum of Transit 1 from the combined result.

Meanwhile, the results for Transits 2 and 3 suggest that little bias has been introduced into the transit spectrum due to the lack of phase coverage in each, while the uncertainties on $\Delta R_p/R_s$ should be larger by $\sim 50\%$ than they would be given full phase coverage.

4.4.7   Correcting for occulted spots

The light curves of Transits 2 and 3 feature prominent spot occultation features that we exclude from the light curve model fit. While doing so allows us to fit the apparent transit depth without simultaneously modeling the spot, it does *not* eliminate the impact of the occulted spot on the resulting transit spectrum. Since the spot is cooler than the rest of the photosphere, it breaks the transit model's assumption that the photosphere is azimuthally

symmetric and amplifies the fitted transit depth in a wavelength-dependent manner (e.g., Deming et al., 2013). In previous works we have referred to this as the transit light source effect (TLSE); here we derive an approximate correction (ignoring limb-darkening) for the TLSE due to a single spot.

Assuming the photosphere to have specific flux $F_{\lambda,\text{phot}}$ and radius R_s, the spot to have specific flux $F_{\lambda,\text{spot}}$ and radius R_{spot}, and the planet to have apparent radius R_p and true radius $R_{p,0}$, then the observed relative transit depth will be

$$\left(\frac{R_p}{R_s}\right)^2 = \frac{F_{\lambda,\text{phot}}R_{p,0}^2}{F_{\lambda,\text{phot}}(R_s^2 - R_{\text{spot}}^2) + F_{\lambda,\text{spot}}R_{\text{spot}}^2} \tag{4.1}$$

By re-arranging this equation, we derive a corrective factor for the transit depth[2]:

$$\epsilon_\lambda \equiv \left(\frac{R_p}{R_{p,0}}\right)^2 = \left[1 - \left(1 - \frac{F_{\lambda,\text{spot}}}{F_{\lambda,\text{phot}}}\right)\left(\frac{R_{\text{spot}}}{R_s}\right)^2\right]^{-1} \tag{4.2}$$

Later, in Section 4.5, we find the feature in Transit 3 to be best described by a model of a single spot with radius $R_{\text{spot}}/R_s = 0.27^{+0.03}_{-0.02}$ and a temperature contrast of $T_s - T_{\text{spot}} = 100 \pm 5$ K. While this feature could in principle be due to a mixture of occulted spots and faculae, our data are insufficient to constrain more complex models, so we use the parameters of the single spot model for the purposes of correcting for the TLSE in this spectrum. We interpolate a PHOENIX (Husser et al., 2013) stellar photosphere model grid onto the values in Table 4.2 to compute $F_{\lambda,\text{phot}}$ and $F_{\lambda,\text{spot}}$ at their respective temperatures, then calculate ϵ_λ from Equation 4.2.

In Figure 4.6 we plot $\epsilon^{1/2}$, which is the corrective factor for the planet *radius*, as well as the original and corrected spectra for Transit 3. The magnitude of the correction from 450 to 900 nm is approximately equal to the binned uncertainty. The corrective factor is listed in Table A.1, binned to match our final spectrum. This corrective factor is not applied to the spectrum of Transit 3 in Figure 4.3, but is applied before creating the combined spectrum in Figure 4.4.

We do not derive a similar correction for Transit 2 because it is unclear (see Section 4.5) whether this feature is best modeled by a small or large spot, and the slope introduced

---

[2]See also Equation 1 of McCullough et al. (2014), Equation 2 of Rackham et al. (2018).
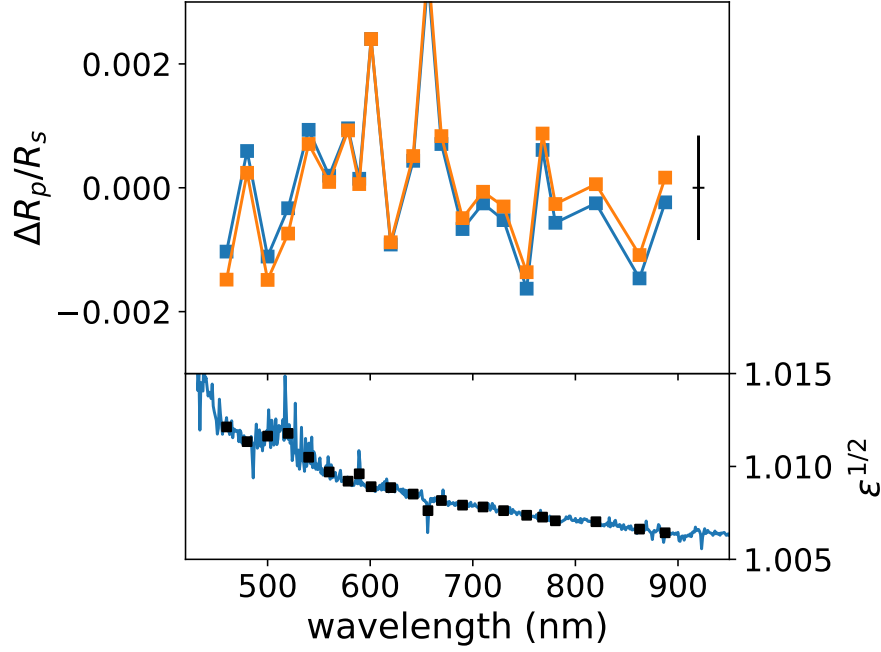
Figure 4.6: (Top) The spectrum of Transit 3 before (blue) and after (orange) correcting for the large occulted spot. Both spectra have been mean-subtracted. The black error bar demonstrates the mean binned uncertainty, while the individual error bars are excluded for visibility. (Bottom) The radius contamination factor due to the spot, binned to 1 nm. The black squares are binned to match the spectrum above.

by the large spot model is six times steeper across the wavelength range than that of the small spot. Furthermore, the large spot model would only introduce a $\sim 0.3\sigma$ increase in the measured radius from the reddest to bluest bins.

Finally, it is worth noting that this effect is just as prominent for spots outside of the transit chord. Equation 4.2 can be further generalized by replacing $(R_{\mathrm{spot}}/R_s)^2$ with F_{het}, the areal spot covering fraction of the photosphere, and using an average spot temperature to compute $F_{\lambda,\mathrm{spot}}$.

4.5 Starspots

In Figure 4.2 we have identified spot-crossing features in the light curves of Transits 2 and 3, and we exclude these data points from our analysis of the transmission spectrum. Such features are common in transit light curves, and have previously been found in tran-

sit observations of WASP-4b (Sanchis-Ojeda et al., 2011). The shape, central time, and magnitude of the feature carry information about the size, position, and contrast (or effective temperature) of the occulted spot, which are otherwise difficult to constrain. Transit spectroscopy allows us to measure the spot contrast at multiple wavelengths, permitting a more precise measurement of the spot's effective temperature than is available from wavelength-integrated light curves.

The measurement of spot properties is worthwhile for a variety of reasons. In the context of this paper, knowing the properties of an occulted spot allows us to correct for its effect on the spectrum of the transit during which it is observed. More generally, observations of multiple spots across an extended period of time permit a better understanding of the makeup of the photospheres of other stars.

4.5.1 Spot modeling

We use SPOTROD (Béky et al., 2014) to produce light curves of transits with one or two spot-crossing events. In addition to the usual transit parameters, SPOTROD models the position and radius (in units of stellar radii) as well as the average spot-to-stellar flux contrast for each spot. We employ PyMultiNest (Buchner et al., 2014) to sample the parameter space and to calculate the Bayesian log-evidences $\ln(Z)$ through which we can compare the different models, as discussed in Section 4.6.1. This MultiNest implementation[3] of SPOTROD has also been used to study spots observed during transits of WASP-19b (Espinoza et al., 2019).

We use a second-order polynomial to detrend the white light curve from each night, then fit the combined spot and star model. We limit the uniform prior on spot size to $R_{\mathrm{spot}}/R_s < 0.08$ or $0 < R_{\mathrm{spot}}/R_s < 1$ to probe for small and large spots. Using the effective temperature in Table 4.2 as a prior, the code then constructs and fits models for the flux contrast as a function of wavelength.

The optimal model parameters for each spot feature are presented in Table 4.4. The precision of the Transit 2 data only permits us to fit the spot feature and spectrum in a single, white-light bin. For this night, we find that our data are consistent with one- and

---

[3]https://github.com/nespinoza/spotnest

two-spot solutions, as well as large and small spots. We opt for the simpler single-spot models and present the fit for both a single large and single cool spot. The data for Transit 3 favor a single large spot versus multiple or smaller spots ($\Delta \ln(Z) > 2$). For this transit we are able to fit the spot spectrum to the same bins as our transmission spectrum, as shown in Figure 4.7.

The large spot model for Transit 3 seems inconsistent with smaller individual spots observed on the Sun; nevertheless, such wide features have been observed in transits before (e.g., Espinoza et al., 2019). While our evidences do not favor a two-spot model, it is possible that the large, low-contrast spot is in fact a more complex active region consisting of multiple spots and/or faculae, as has been observed on the Sun. Regardless of the actual structure of the feature, its average contrast is well-described by a single spot with $T_s - T_{\text{spot}} = 100$ K, and we choose this model to calculate the correction in Section 4.4.7.
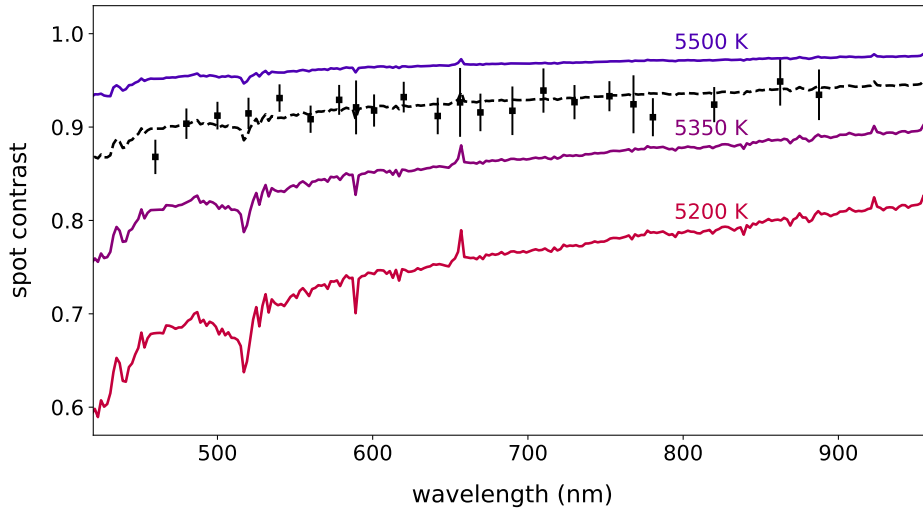
Figure 4.7: Best-fit model for the occulted spot in the light curve of Transit 3. The data are best described by a single large, warm spot with a radius of $0.27 R_s$ and a temperature contrast of 100 K.

Table 4.4.   Best-fit Parameters for the Spot-crossing Events in the Light Curves of
Transits 2 and 3

	R_{spot}/R_s	f_{spot}/f_s	$T_s - T_{\text{spot}}$ (K)	T_{spot} (K)
Transit 2a	$0.05^{+0.01}_{-0.01}$	$0.49^{+0.20}_{-0.28}$	880 ± 360	4670 ± 360
Transit 2b	$0.15^{+0.11}_{-0.07}$	$0.79^{+0.11}_{-0.18}$	410 ± 280	5140 ± 280
Transit 3	$0.27^{+0.03}_{-0.02}$	$0.91^{+0.01}_{-0.01}$	100 ± 5	5442 ± 50

Note. — The data for Transit 2 are consistent with both a small (2a) and large spot (2b). Median and 1σ confidence intervals are reported.

4.5.2   Spot sizes and temperatures

Later, in Section 4.6, we assume a spot temperature in order to model the effects of unocculted spots on the transit spectrum. The temperature we assume reflects small, cool spots resembling those on the Sun, but the results for Transit 3 in Table 4.4 appear to suggest the existence of large, warm features on the photosphere of WASP-4. In short, we do not believe that the spots discovered in our transit light curves are necessarily representative of the spots present elsewhere in the photosphere or below our detection threshold. We offer these arguments as justification:

1. Sanchis-Ojeda et al. (2011) find evidence for a persistent small spot in multiple transit observations of WASP-4b. In this case, they find a spot with radius $R_{\text{spot}}/R_s \approx 0.05$ and temperature contrast $T_s - T_{\text{spot}} \approx 600$ K, suggesting that more Sun-like spots do exist on WASP-4.

2. Cool spots have previously been discovered through Doppler imaging of active G and K dwarfs (e.g., O'Neal et al., 1998, 2004; Strassmeier, 2009). However, this

method of spot detection is subject to biases regarding spot size and temperature.

3. The feature occulted during Transit 3 may indeed have a more complex structure consisting of cool spots and hot faculae that the precision of our data does not allow us to resolve.

4. While the precise radius and temperature of the spot observed during Transit 2 are largely unconstrained, we can say with at least 1σ confidence that the spot is smaller and cooler than the feature observed during Transit 3. Therefore, even our data suggest that small spots may be present in the photosphere of WASP-4.

It is nevertheless possible that the distribution of spot sizes on WASP-4 favors larger and hot spots, in which case the spectral contamination signal would be less severe. However, our use of the contamination model is not meant to accurately account for the actual surface heterogeneity, but rather to demonstrate that this aspect should not be ignored in determining the planet's transmission spectrum.

4.6   Atmospheric retrieval

To interpret the combined transmission spectrum in Figure 4.4, we employ a Bayesian atmospheric retrieval code based on PyMultiNest that has previously been used to study WASP-19b (Espinoza et al., 2019). Following Bétrémieux and Swain (2017) and Heng and Kitzmann (2017), our atmosphere model includes two regions: an optically thick base region (which could also be interpreted as a cloud layer) with radius $(R_p/R_s)_0$ and reference pressure P_0, and an isothermal optically thin region above with temperature T. The components of the optically thin region may include either or both of (i) a set of atomic and molecular species and (ii) a scattering haze defined by a cross-section power law $\sigma = a\sigma_0 \left(\lambda/\lambda_0\right)^{\gamma}$, where $\sigma_0 = 5.31 \times 10^{-27}$ cm^2 is the Rayleigh scattering cross-section of H$_2$ at the reference wavelength $\lambda_0 = 350$ nm (MacDonald and Madhusudhan, 2017), and a is a dimensionless enhancement factor. We constrain γ to be between 0 (uniform opacity) and -4 (Rayleigh scattering), which should span the range of physical scattering processes.

As we have previously argued in Rackham et al. (2017, 2018), it is important that analyses of transmission spectra account for the transit light source effect (TLSE) that can be introduced by heterogeneous features (spots and faculae) on the stellar surface. We incorporate a three-parameter model for the stellar photosphere into our retrieval code, fitting this model simultaneously with that of the planet's atmosphere. Section 4.6.2 discusses this model in more detail.

The parameters and priors for each component of the model are listed in Table 4.5. We consider eight models for the atmosphere, including all possible combinations of a cloud deck, a scattering haze, and Na and/or K absorption signals. Each of these is fitted independently or alongside a photosphere model (+1 free parameter, F_{het}) for a total of sixteen models.

4.6.1 Model comparison

To compare between our models, we compute their relative Bayes factors. A thorough introduction to the use of Bayes factors in astronomy is given by Trotta (2008), and their use in transit spectroscopy is well-established (e.g., Benneke and Seager, 2013). Here, we provide a brief overview.

Given a model M with parameters θ, the optimal values of θ to describe data values x are those that maximize Bayes' equation:

$$P(\theta|x,M) = \frac{P(x|\theta,M)P(\theta|M)}{P(x|M)}$$

where $P(x|\theta,M)$ is the likelihood function and $P(\theta|M)$ is the prior distribution of θ. Most Bayesian model fitting algorithms seek to maximize the likelihood function, but the maximum likelihoods of two different models cannot be directly compared.

To compare two models, we first compute their Bayesian evidences, defined as:

$$Z_M \equiv \int P(x|\theta,M)P(\theta|M)d\theta$$

The evidence for a model is high if the data are well-described by a large region of the prior parameter space. The Bayes factor between two models is defined as the ratio of their evidences:

$$\ln(B) \equiv \ln(Z_2) - \ln(Z_1)$$

In general, model 2 is weakly favored over model 1 if $0 < \ln(B) < 2$, and strongly favored if $\ln(B) > 5$. If $\ln(B) < 0$ then model 2 is disfavored to model 1. Complex models will be favored over simpler models only if the evidence justifies the additional parameters. However, a preference towards simpler models does not necessarily mean that the complex models are incorrect, but rather that they are not justified by the data at hand.

Our retrieval algorithm computes the Bayesian evidence parameter for each of the fitted models. In the following sections, we cite the Bayes factor $\ln(B)$ for each model with respect to the model for a uniform opacity atmosphere with no contamination from the photosphere (i.e., a flat line). In general, we find that the more complex models have $\ln(B) < 0$, meaning that they are *dis*favored versus a flat line.

Table 4.5.   Parameters for Our Combined Photosphere and Atmosphere Models

Model component	Parameter	Units	Description	Prior distribution
Base	$(R_p/R_s)_0$	-	Radius corresponding to the top of the cloud layer or $\tau \gg 1$	Uniform(0.1,0.2)
	$P_0{}^\dagger$	bar	Reference pressure at $(R_p/R_s)_0$	Log-uniform(10^{-6},10^6)
	$T^\dagger$	K	Average temperature of the optically-thin region	Uniform(1000,2400)
Atomic features	X	-	Mixing ratio of species X	Log-uniform(10^{-30},1)
Haze	a	-	Amplitude of the haze cross-section power law	Log-uniform(10^{-20},10^{10})
	γ	-	Index of the haze cross-section power law	Uniform(-4,0)
Stellar photosphere	T_{occ}	K	Average temperature of the transit chord[‡]	Fixed(5540)
	T_{het}	K	Average temperature of the heterogeneous surface features	Fixed(4200)
	F_{het}	-	Fraction of the unocculted photosphere covered by spots	Uniform(0,1)

Note. — Not all parameters are incorporated in all models; rather, Bayesian log-evidences are used to compare separate models.

[†]The temperature and reference pressure are not modeled in the case of a uniform opacity atmosphere.

[‡]This excludes any occulted features that can be identified in the light curve.

4.6.2 Stellar heterogeneity

Assessing the level of TLSE contamination is critical to correctly interpreting high-precision transmission spectra of transiting exoplanets (Apai et al., 2018b). This is particularly important for investigating features that could have a planetary or stellar origin, such the He 10,833 Å line (Spake et al., 2018) or the Na and K alkali lines (Cauley et al., 2018). We have previously demonstrated this effect in the spectrum of GJ 1214b (Rackham et al., 2017), which has a featureless near-IR spectrum but a visible spectrum that is apparently offset below the near-IR transit depths. In this study, we found that for reasonable assumptions about the star's activity, the planetary transmission spectrum of GJ 1214b can be shown to be flat in the visible as well. We also detect the signature of stellar activity in one of our *Magellan*/IMACS transit data sets for WASP-19b (Espinoza et al., 2019).

WASP-4 is of slightly later spectral type than the Sun, and multiple spot-crossing events have been observed during previous transit observations of WASP-4b (Sanchis-Ojeda et al., 2011) as well as in two of our data sets. To assess the photometric variability, we retrieve four years of photometric monitoring data from the ASAS-SN online database[4] (Shappee et al., 2014; Kochanek et al., 2017). We remove outliers with a 5σ cut and create a Lomb-Scargle periodogram to search for periodic signals between 1 and 100 days, calculating the false alarm probability as in Section 4.4.5. As shown in Figure 4.9, we find the largest peak at 22.4 days, and while the significance of the detection is only moderate, it corresponds well with the rotational period of 22.2 days that Sanchis-Ojeda et al. (2011) measure from consecutive spot-crossing events. This is to be expected if the photometric variability is dominated by the rotation of spots and faculae in and out of view. Modeling the variability as a sine curve and assuming a period of 22.4 days, we perform a least-squares fit to the phase folded data to find a peak-to-trough variability amplitude of ~ 6 mmag (0.6%). These results suggest that the surface of WASP-4 is moderately heterogeneous, so any analysis of the transmission spectrum of WASP-4b that ignores stellar contamination may be overly optimistic. Furthermore, since the spot

---

[4]https://asas-sn.osu.edu/
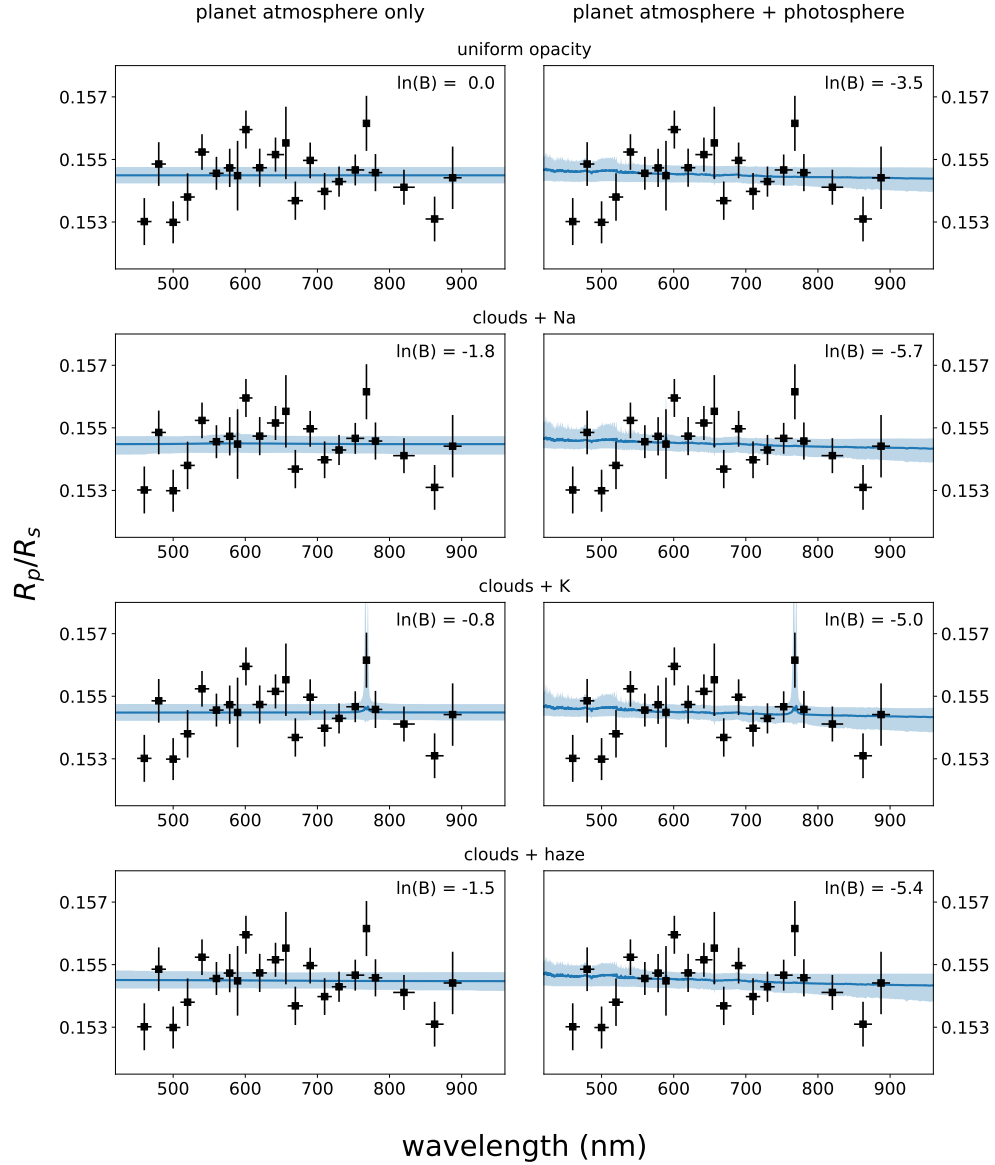
Figure 4.8: A subset of fitted models for the observed transmission spectrum of WASP-4b; four separate models for the planetary atmosphere are presented alone (left) or combined with a model for the heterogeneous stellar photosphere (right). The atmosphere models, described in more detail in Section 4.6, include (top) a uniform opacity atmosphere, (middle) a cloud deck and *Na* absorption feature, a cloud deck and *K* absorption feature, and (bottom) a cloud deck and scattering haze. The shaded region represents the 95% confidence interval around the mean model fit. Bayes factors are given for each model relative to the flat line model (upper left). A constant offset of +0.1545 has been added to the combined spectrum.
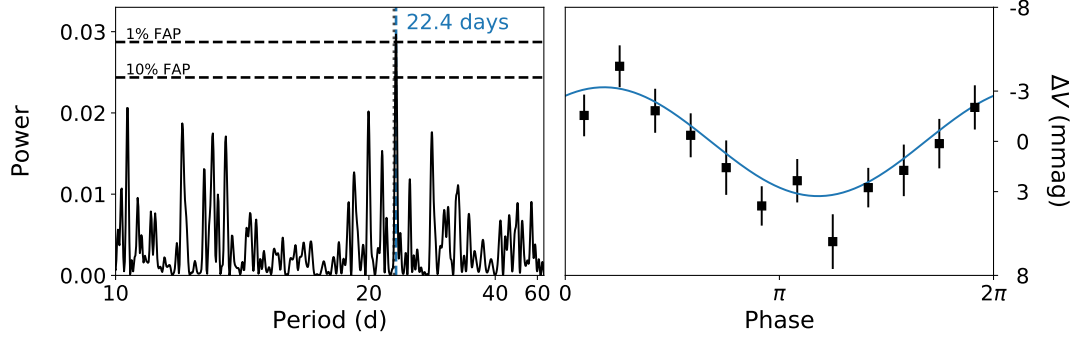
Figure 4.9: (Left) Lomb-Scargle periodogram for the *V*-band magnitude of WASP-4 from ASAS-SN photometry. The largest peak, marked with a blue line, corresponds to the previously measured rotation period, marked with a black dotted line. The horizontal lines denote the false alarm probability calculated through the bootstrapping method. (Right) Fitted model for the phase-folded light curve, assuming a period of 22.4 days. The data are binned for visibility.

covering fraction can not be determined from the variability alone (Rackham et al., 2018), the photosphere heterogeneity should be modeled alongside the atmosphere.

In the retrieval code described above, we characterize the heterogeneity with three parameters: T_{occ}, the effective temperature of the transit chord, T_{het}, the mean effective temperature of the heterogeneous features, and F_{het}, the fraction of the unoccluded photosphere that is covered in spots. We then compute the effect on the observed transit depth following Section 4.4.7 and Equation 4.2, but replacing the area ratio $(R_{spot}/R_s)^2$ with a covering fraction F_{het}.

We test more complex parameterizations that include multiple types of unoccluded heterogeneities (e.g., Zhang et al., 2018), but determine from the evidences that they are not warranted by the data.

Given the quality of our data, we find it necessary to fix some parameters of the heterogeneity model to reasonable values. In Table 4.5 we fix the spot temperature to match typical spots on the Sun (e.g., Solanki, 2003, and references therein), and the transit chord temperature to the measured effective temperature of the photosphere.

F_{het} is allowed to vary over all possible values, and serves as a simple measure of the level of stellar contamination in the transmission spectrum. However, we note that F_{het}

represents an *enhancement* over the effect of the large occulted spot in Transit 3, which has already been corrected for in Section 4.4.7.

4.6.3 Retrieval results

In Figure 4.8 we present four of our atmosphere models with and without photosphere contamination. In Table 4.6 we list the Bayes factors for all sixteen of the models tested.

4.6.3.1 Favored models

The evidences favor a uniform opacity model for the atmosphere, with other atmosphere models being disfavored by $\ln(B) = -1$ to -2. Even when the haze is included, the amplitude a of the haze opacity is small; the posterior distributions do not converge well enough from the log-uniform prior to offer a meaningful upper limit, but the mode of the distribution for a lies between 10^{-18} and 10^{-12}, which is lower than in the prior distribution.

Atmosphere models including Na and/or K are disfavored versus uniform opacity models by $\ln(B) = -1$ to -2. Even though the narrow K bin is elevated above the continuum by $\sim 1\sigma$, ultimately the presence of potassium is not warranted by the data. We note that we cannot place upper limits on the abundances of Na or K, as these are degenerate with the reference pressure.

The models that include a heterogeneous photosphere are strongly disfavored by $\ln(B) = -5$ to -6 versus their counterparts with a homogeneous photosphere. While it is quite likely that $F_{\text{het}} > 0$, the spectrum *alone* does not reveal evidence for it beyond what we have already corrected for in Section 4.4.7. The effect of even a low spot covering fraction ($< 3\%$) is apparent in Figure 4.8, and is degenerate with the effect of a scattering haze. For this reason, we recommend the joint modeling of the photosphere and atmosphere in future analyses of optical transmission spectra, and the use of Bayesian evidences or likelihood parameters for model selection (see also Pinhas et al., 2018).

4.6.3.2 Constraints on spot covering fraction

The parameters in the clear and hazy atmosphere models do not converge enough from their priors to offer meaningful constraints. However, in the special case of a uniform opacity atmosphere with stellar contamination, we can place a 95% upper limit on the spot covering fraction, $F_{\text{het}} < 3.4\%$. We caution that this is under the assumption of 4200 K spots and does not factor in the likely diversity of spot and faculae temperatures on the stellar surface. Instead, it should be taken as a limit on the net spectral effect of contamination.

4.6.4 Correcting vs. fitting the contamination in Transit 3

In Section 4.4.7 we detail a method for correcting the contamination due to the cool feature that was occulted by the planet during Transit 3. In this retrieval, however, we also include a varying parameter F_{het} and fixed parameter T_{het} to characterize the heterogeneity of the photosphere. We offer the following argument to justify our decision to visit the heterogeneity twice.

First, while the detailed structure of the occulted feature from Transit 3 is unclear, it is well-described by a circular spot model with a constant temperature, and this is the same model that we use to correct for its contamination. By excluding this correction, we would not be leveraging all of the available information from our light curve. However, F_{het} must still be non-zero to describe the remaining unocculted features.

Second, we demonstrate that this correction is accounted for during the atmospheric retrieval. As a test, we fit the model of a uniform opacity atmosphere with a heterogeneous photosphere (Figure 4.8, top-right), this time fixing $T_{het} = 5442$ K to match the temperature of the occulted feature. When we fit the uncorrected combined spectrum, we find that the median of the posterior distribution for F_{het} is larger than for the corrected combined spectrum. The difference is consistent with a single spot of size $R_{spot}/R_s \approx 0.22$. Similarly, when the retrieval is applied to the corrected and uncorrected spectra from Transit 3 only, the difference in the median value of F_{het} is consistent with a single spot of size $R_{spot}/R_s \approx 0.33$. Both of these are similar to the spot size assumed in the correction,

Table 4.6.   Bayes Factors for the Full Suite of Atmosphere Models

Model	$\ln(B)_A$	$\ln(B)_{A+P}$
uniform opacity	0.0	-3.5
clouds + Na	-1.8	-5.7
clouds + K	-0.8	-5.0
clouds + Na,K	-1.0	-5.4
clouds + haze	-1.5	-5.4
clouds + haze + Na	-1.9	-6.2
clouds + haze + K	-1.9	-5.4
clouds + haze + Na,K	-1.9	-6.2

Note. — The two columns repre-
sent atmosphere-only (A) and combined
atmosphere-photosphere (A+P) models. $\ln(B)$
is calculated relative to the flat line model;
since $\ln(B) < 0$ for the more complex models,
they are all disfavored versus a flat line. Some
of these model fits are shown in Figure 4.8.

$R_{spot}/R_s = 0.27^{+0.03}_{-0.02}$. This indicates that the effect of the correction is carried forward into the retrieval; we are not over-correcting for the effect of the spot.

4.7 Comparison to published results

4.7.1 *Gemini*/GMOS

Huitson et al. (2017) (hereafter H17) have previously published an optical transmission spectrum of WASP-4b from four combined transit observations with *Gemini*/GMOS, and find that their data favor a uniform opacity model. In this section, we assess the agreement of their results with our data from IMACS. In Figure 4.10 we compare the red and blue GMOS spectra with the combined spectrum from this work, and find that the slopes of our spectra are generally in agreement.

4.7.1.1 Absorption features

The H17 data reveal tentative evidence for Na absorption at 589 nm. Our data reveal no evidence for such a feature, but we concede that the quality of our data in the blue is not sufficient to definitively rule out an atomic feature of small optical depth.

H17 exclude bins in their spectra that are coincident with terrestrial telluric O_2 absorption, finding that their correction for differential atmospheric refraction was not reliable in this wavelength range. Magellan is equipped with an ADC so our data do not require this correction (see Section 4.3.4). As a result, we are able to test for a K I feature in the same region, and find little evidence as discussed in Section 4.6.

4.7.2 *Magellan*/IMACS

May et al. (2018) (hereafter M18) used *Magellan*/IMACS to study this target with the same grism as we used for Transits 1-3, and they also find no evidence for features in the transit spectrum. Figure 4.10 demonstrates that our results are in agreement.

4.7.2.1 Lack of a spot crossing event

The transit observed by M18 occurred on 19 August 2015, approximately 5.4 days after our Transit 3, though their light curves show no evidence for a large occulted photosphere feature. This is not unexpected: the planet's period is prograde and nearly equatorial, and the stellar rotation period is ~ 22.2 days (Triaud et al., 2010; Sanchis-Ojeda et al., 2011). This implies that the feature observed during Transit 3, which was occulted in the second half of the transit, would have rotated ~ 90 degrees and off the projected stellar disk within the period of time between the transits.

4.7.3 Combined analysis

We combine the data from H17 and M18 with our own by subtracting the weighted mean from each spectrum, then repeat the retrieval analysis of Section 4.6. Table 4.7 displays the Bayes factor relative to a flat line for each for the 2x8 models. As before, we find that most atmosphere models are disfavored versus a uniform opacity model by $\ln(B) = -1$ to -2, but models including Na are slightly less disfavored than in Table 4.6. This indicates that there is *more* evidence for Na absorption in the combined data set than in ours alone, but still not enough to justify its inclusion.

Models that include the photosphere are disfavored by $\ln(B) \sim -5.5$. In the case of a uniform opacity atmosphere, the 95% upper limit on the spot covering fraction is $F_{\rm het} < 1.8\%$, which is smaller than the value we report in Section 4.6.3.2. However, we caution that the red and blue GMOS spectra were observed during separate transits, which may mask the slope introduced by TLSE contamination.

4.8 Featureless Atmosphere

Our transit spectra, in combination with the spectra by H17 and M18, suggest that WASP-4b displays a mostly featureless (uniform opacity) spectrum without a strong optical spectral slope or alkaline absorption (Na I or K I lines). Although more observations will be required to place tighter constraints on the optical spectrum to verify this, in the following we will explore what a uniform opacity spectrum would suggest for WASP-4b.
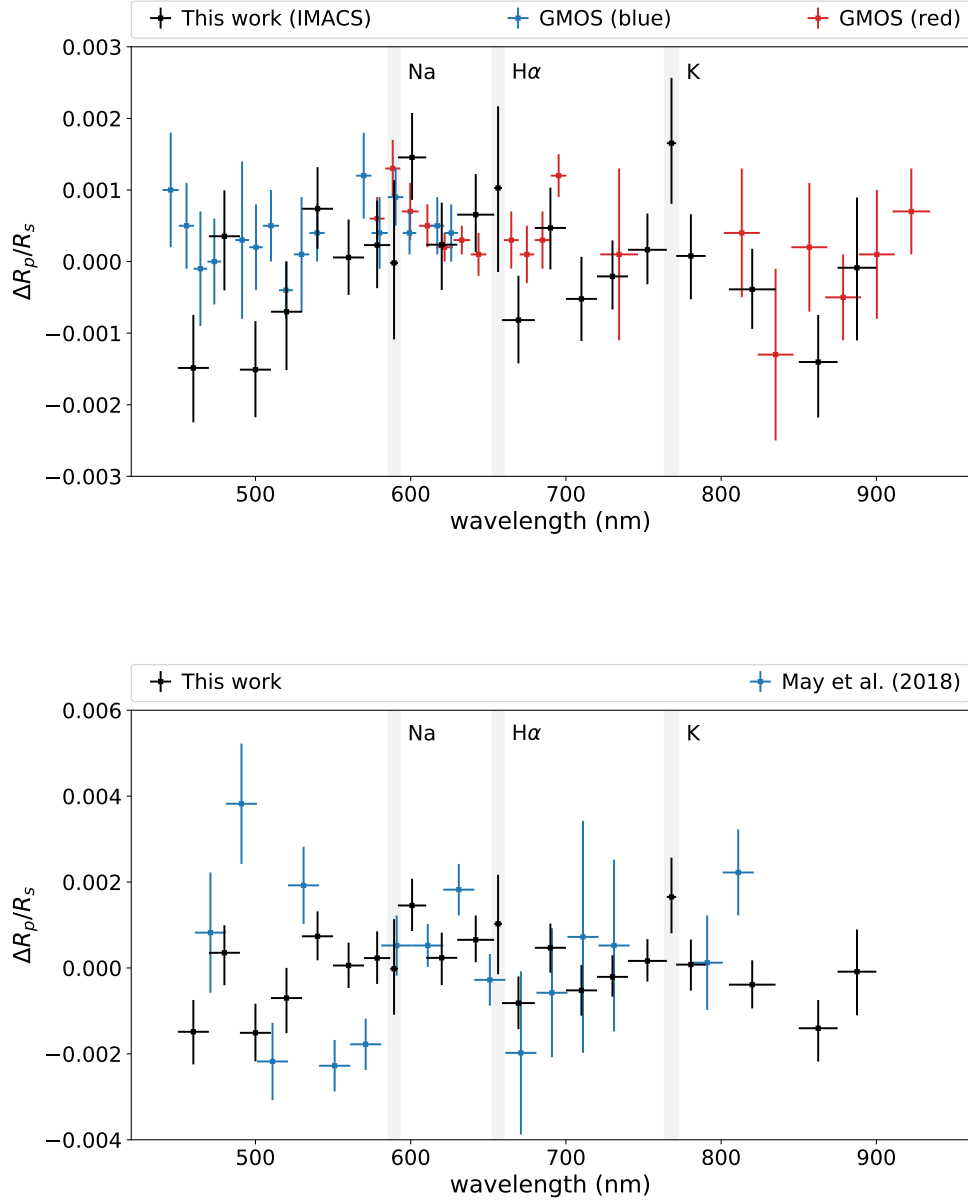
Figure 4.10: A comparison of our combined transmission spectrum to the spectrum of (top) Huitson et al. (2017) (Table 4) as observed with *Gemini*/GMOS and (bottom) May et al. (2018) (Table 5) as observed with IMACS. The wavelengths of potential atomic features are highlighted. The weighted mean is subtracted from each of the three data sets.

Table 4.7. Bayes Factors for the Full Suite of Atmosphere Models, Including Data from Other Analyses

Model	$\ln(B)_A$	$\ln(B)_{A+P}$
uniform opacity	0.0	-3.8
clouds + Na	-0.2	-5.5
clouds + K	-1.3	-5.7
clouds + Na,K	-0.7	-4.9
clouds + haze	-1.7	-6.4
clouds + haze + Na	-1.2	-6.0
clouds + haze + K	-1.6	-6.3
clouds + haze + Na,K	-1.6	-5.9

.

Note. — Includes data from our analysis as well as H17 and M18. Column definitions are similar to Table 4.6

Cloud-free, broadly solar-composition atmospheres are predicted to display strong absorption at visible wavelengths by alkali metal atoms (Na I or K I doublets, Seager et al. 2000; Sudarsky et al. 2000; Seager and Sasselov 2000). These prominent (deep and broad) features have been observed in the spectra of multiple transiting hot Jupiters (e.g., Charbonneau et al., 2002). The depths of the truncated alkali features are often used as proxy for the atmospheric pressures probed (e.g., Sing et al., 2015). The emerging evidence argues that transit sightlines are often limited by cloud decks, which then lead to truncated alkali line cores or, for very low-pressure particles, even the absence of detectable alkali absorption.

Another marked deviation from a flat (featureless, zero-slope) visible spectrum would arise in a clear atmosphere from Rayleigh scattering of the starlight by molecules or very small particles. Rayleigh scattering is more efficient at smaller wavelengths, resulting in apparently larger planet diameters (i.e., a lower pressure level $\tau = 1$ surface) at shorter wavelengths (e.g., Pont et al., 2008). The actual wavelength-dependence of the Rayleigh-scattering in a given transiting exoplanet atmosphere will depend on the particle size distribution in its upper atmosphere and can vary by orders of magnitude (e.g., Heng and Kitzmann 2017). However, visual-wavelength slopes in transiting exoplanets may also be introduced by the transit light source effect described in Section 4.4.7; given this consideration, the simultaneous presence of visual slopes and alkali line absorption are considered to be the most robust indicators of clear (particle-free) upper atmospheres.

Therefore, with multiple transit spectra suggesting the lack of alkali absorption *and* the lack of a strong visual slope for WASP-4b, it is important to consider the possibility that WASP-4b's upper atmosphere is not clear but contains particles at high-altitude. Given this strong possibility, in the following we explore the possible nature of these particles and the mechanisms that may inject particles into the upper atmosphere. Based on *Spitzer*/IRAC eclipse (dayside emission) measurements in the [3.6] and [4.6] filters, Beerer et al. (2011) found that the best-fit blackbody temperature of WASP-4b's dayside is 1,700 K. This temperature is close to the radiative equilibrium estimate of 1,650 K (assuming zero albedo) and significantly lower than 2,000 K, the temperature that would correspond to a zero-albedo, dayside-emission-only case.

The dayside temperatures of WASP-4b are hot enough to vaporize less refractory grains and – in a smaller fraction of the hemisphere centered on the sub-stellar point – it is likely hot enough to vaporize metal-oxide and silicate grains. Cloud formation through the evaporation and condensation of refractory grains is common and relatively well-studied in non-irradiated brown dwarfs (e.g., Apai et al., 2013; Buenzli et al., 2014; Metchev et al., 2015) and directly imaged exoplanets (Biller et al., 2015; Zhou et al., 2016) of similar temperatures. Equilibrium condensation models coupled to global circulation models show that refractory species (e.g., $CaTiO_3$, $MgSiO_3$, MnS, Na_2S) will also form clouds in the upper atmospheres (typically 10–100 mbar) of hot Jupiters (e.g., Parmentier et al., 2016; Showman et al., 2009; Kataria et al., 2016). Therefore, our observations suggesting the lack of evidence for a clear atmosphere are fully consistent with the general expectations set by observations and models of brown dwarf and hot Jupiter atmospheres of similar temperatures.

4.9 Conclusions

We have extracted a combined optical transmission spectrum from three transit observations of WASP-4b using *Magellan*/IMACS, and use a MultiNest-based retrieval code to test different atmospheric models for the data. We find that a uniform opacity model for the atmosphere is weakly favored over alternatives with hazes, Na, and/or K. In particular, no meaningful evidence for potassium is found despite the elevated radius in the narrow bin centered on the K doublet. Our results, in addition with those of Huitson et al. (2017) and May et al. (2018), suggest that no strong signals exist in the optical transit spectrum. Nevertheless, higher quality data may yet reveal evidence for scattering or atomic absorption; for example, Huitson et al. (2017) find inconclusive evidence for Na absorption.

We are also able to fit the size and contrast of the star spots occulted by the planet during Transits 2 and 3. Assuming a single spot model, the quality of the data from Transit 3 allows us to tightly constrain the spot size and contrast, which suggest a spot that is much larger and warmer than is typical for spots on the Sun. More complex models that include several spots and faculae could be consistent with the data as well, but are not

justified by the evidences. We use this spot model to correct for the wavelength-dependent effect on the transmission spectrum from Transit 3 before averaging the individual nights' spectra.

Further space-based or larger aperture ground-based observations should be conducted to search for low amplitude signatures of scattering or absorption. However, we note that the presence and strength of a scattering haze is particularly degenerate with the presence of spots and faculae on the star. Since WASP-4 is known to be variable, the stellar photosphere and planet atmosphere should be modeled simultaneously in future analyses of this planet's transmission spectrum.

4.9.1 Acknowledgements

CHAPTER 5

Testing Earth-like Atmospheric Evolution on Exo-Earths through Oxygen Absorption: Required Sample Sizes and the Advantage of Age-based Target Selection

This chapter has been published as Bixel and Apai (2020b) and is reproduced here with permission.

Abstract

Life has had a dramatic impact on the composition of Earth's atmosphere over time, which suggests that statistical studies of other inhabited planets' atmospheres could reveal how they co-evolve with life. While many evolutionary pathways are possible for inhabited worlds, a possible starting hypothesis is that most of them evolve similarly to Earth, which we propose could lead to a positive "age-oxygen correlation" between the ages of inhabited planets and the fraction that have oxygen-rich atmospheres. We demonstrate that next-generation space observatories currently under consideration could test this hypothesis, but only if the stellar age distribution of the target sample is carefully considered. We explore three possible parameterizations of the age-oxygen correlation, finding that they yield similar results. Finally, we examine how abiotic oxygen sources could affect the results, and discuss how measuring the age-dependence of oxygen could shed light on whether it is a reliable biosignature. Future efforts can expand upon this groundwork by incorporating detailed models of the redox balance of terrestrial planets and its dependence on stellar and planetary properties.

5.1 Introduction

The coming decades promise exciting developments in the search for life beyond Earth, with multiple groups proposing the construction of novel space observatories that could discover and characterize several potentially Earth-like planets orbiting nearby stars (e.g., Gaudi et al., 2020; The LUVOIR Team, 2019; Apai et al., 2019; Meixner et al., 2019; Staguhn et al., 2019). By discovering biosignature molecules in these planets' atmospheres (e.g., Schwieterman et al., 2018), such observatories would enable the first constraints on the frequency of life in the universe and comparative studies of the properties of inhabited worlds.

Molecular oxygen (O_2), and its photochemical byproduct ozone (O_3), have been discussed as promising biosignatures for such missions, as O_2 has a short lifetime in the Earth's atmosphere and is replenished almost entirely by photosynthetic life (e.g., Owen, 1980; Des Marais et al., 2002). O_2 would make for an even stronger biosignature if it were found in the presence of reduced gasses (such as methane) that would quickly eliminate it in the absence of a strong oxygen source (e.g., Lovelock, 1965; Meadows et al., 2018b). Oxygenic photosynthesis makes use of carbon dioxide, water, and light, which have been accessible on Earth throughout its history - suggesting that many extraterrestrial ecosystems may have converged on the same mechanism (e.g., Meadows, 2017).

However, the presence of oxygen in Earth's atmosphere has evolved over time, with the planet having an anoxic atmosphere for approximately the first half of its history. During the Hadean and Archean eras, the abundance of O_2 was no more than 10^{-6} times its present atmospheric level (PAL) (Zahnle et al., 2006; Catling and Zahnle, 2020). Then, during the "Great Oxidation Event" (hereafter GOE) circa 2.4-2.1 Gya, the concentration of O_2 dramatically increased to between 10^{-4} to 10^{-1} PAL, and would later increase again (circa 600 Mya) to reach its modern abundance (Lyons et al., 2014).

The precise causes and timing of the GOE are a matter of ongoing research - for a thorough review, see Lyons et al. (2014). While the evolutionary development of oxygenic photosynthesis was a prerequisite for the GOE to occur, the two were not necessarily coeval; in fact, the evidence suggests a delay - perhaps hundreds of Myr long - between

the appearance of the first organisms to produce oxygen and its eventual accumulation in the atmosphere (e.g., Brocks et al., 1999; Anbar et al., 2007; Kendall et al., 2010; Kurzweil et al., 2013; Planavsky et al., 2014). Regardless of its causes, the GOE counts among the most dramatic changes to Earth's atmosphere in geological history, and was dependent on the existence of life. Since Earth's atmosphere was anoxic for about half of its history, some authors have considered how pre-GOE Earth analogs might appear in reflected or transmitted light, and how the presence of life on such worlds could be inferred in the absence of oxygen (e.g., Pilcher, 2003; Domagal-Goldman et al., 2011; Seager et al., 2013; Arney et al., 2016, 2018; Krissansen-Totton et al., 2018).

To date, most studies of oxygen and other potential biosignatures have focused on how life could affect individual planets. However, due to the challenging nature of characterizing terrestrial planets, future space telescopes may provide only limited information about the atmospheres and fundamental parameters (e.g. bulk composition) of many individual planets. Even still, important information will be enclosed in the overall population of planets studied, and trends between their properties can be tested against the predictions of models for terrestrial planet evolution (e.g., Bean et al., 2017; Checlair et al., 2019). In such cases, sample sizes will be a limiting factor on the complexity of models that can be tested.

Since Earth's biosphere and atmosphere have co-evolved over time, the exciting possibility exists that by studying several inhabited planets spanning a range of ages, we could test for shared trends in the co-evolution of their atmospheres and biospheres, uncovering common patterns that govern the evolution of life in the universe. For example, if future space missions are able to detect the presence of O_2 or O_3 in the atmospheres of several potentially habitable worlds, this would allow them to constrain the frequency of oxygen-bearing planets as a function of age. We propose that this measurement could be used to test whether the atmospheric evolution of other inhabited worlds resembles that of Earth (i.e. the null hypothesis). By this, we mean that they start with anoxic atmospheres that eventually become oxygen-rich - although the time required for oxygenation will likely vary between planets (Catling et al., 2005). If they evolve like Earth, then the fraction of inhabited planets with oxygenated atmospheres should increase with age, with

older planets being more likely to have undergone a GOE-like event. If such a trend were discovered, it would strongly suggest that Earth-like atmospheric evolution is typical for inhabited planets, and would by extension strengthen the case for O_2 as a biosignature, as we can think of no plausible abiotic explanation for this trend.

In this paper, we estimate how many potentially habitable planets a future telescopic biosignature survey must characterize to detect a positive "age-oxygen correlation" – and thus test whether Earth's atmospheric evolution is typical. We present our results as a function of the actual occurrence rate of life on potentially habitable worlds, and investigate the optimal target stellar age distribution for testing our hypothesis. Finally, we estimate the noise that would be introduced by strong abiotic sources for O_2, and discuss how a statistical sample of planets with oxygenated atmospheres could be used to verify O_2 as a biosignature.

This study does not attempt to model in detail the many possible factors affecting the rate at which inhabited planets acquire oxygen, which we discuss qualitatively in Section 5.4.2. Rather, we use Earth's evolutionary history as the practical template for an initial estimate of the sample size required to begin studying their oxygen evolution. Our results should be interpreted with this caveat in mind, and we encourage future studies to build off our approach by incorporating the effects of diverse planetary parameters on the redox balance of other habitable worlds.

5.2   Methods

Here, we consider only potentially habitable planets or "exo-Earth candidates" (EECs). We use these terms interchangeably to refer to planets that are comparable in size to Earth and have orbits within the liquid water habitable zone, and that therefore could sustain habitable surface conditions (e.g., Kasting et al., 1993; Kopparapu et al., 2013, 2014). Note, however, that our results are agnostic to the exact range of sizes and orbits considered to be potentially habitable, except when we compare them to the predicted discovery yields for future space-based biosignature surveys. For the reasons argued above, we also assume that inhabited planets demonstrate a positive correlation between their ages and

the fraction that have oxygen in their atmospheres, which we hereafter refer to as the "age-oxygen correlation".

Our basic methodology is as follows: we generate a number N EECs with randomly assigned ages t, a fraction f_{life} of which are assumed inhabited. We then assume that some age-dependent fraction $f_{O_2}(t)$ of inhabited planets have undergone a GOE-like event and therefore have detectable amounts of O_2 *or* O_3 in their atmospheres. Assuming that a future space-based survey has discovered and spectroscopically characterized the entire sample, we apply a statistical test to calculate the confidence with which the age-oxygen correlation could be discovered as a function of the number of planets observed. Finally, we average the results of this test across 10^4 random samples[1] for each cell in a two-dimensional grid of values for N and f_{life}.[2]

5.2.1 Fraction of inhabited planets with O_2

We consider three functions to describe the fraction $f_{O_2}(t)$ of inhabited planets that have detectable O_2 or O_3 in their atmospheres as a function of their age t (or, approximately, the age of their host star). We consider the stellar ages to be determined precisely, and discuss the feasibility of this assumption in Section 5.4.6.

The functions, plotted in Figure 5.1, are as follows: in the first case, $f_{O_2}(t)$ increases exponentially over an e-folding timescale of 3.2 Gyr, so that the typical planet reaches its GOE at the same epoch as Earth's ($t \sim 2.2$ Gyr). In the second case, all planets evolve identically, encountering their GOEs at the same point in time as Earth did. The third case assumes the same functional form as the first, but with a longer timescale of 10 Gyr - in this case, most planets encounter their GOEs much later than Earth did. We discuss the motivation behind these functional forms in Section 5.4.5.

Finally, in each case we only distinguish between oxygenated and anoxic atmospheres; the mixing ratio of O_2 is not modelled. Generally speaking, the presence or absence of O_2 will be much easier to determine than its precise mixing ratio. Further-

---

[1]In some samples none of the planets have O_2, so the *p*-value is undefined and we discard it. This generally only occurs for low values of N and f_{life} and does not significantly impact our results.

[2]The code used to generate the grid of p-values and Figures 5.1-5.3 can be found here.

more, while O_2 may not be directly detectable for planets with only a small biogenic abundance – analogous to Earth during the Proterozoic era – its presence might still be inferred through that of its strongly absorbing byproduct O_3 (e.g., Angel et al., 1986; Des Marais et al., 2002; Segura et al., 2003; Reinhard et al., 2017).

5.2.2 Abiotic O_2

Several authors have suggested scenarios in which a rocky planet in or near the habitable zone could attain a detectable amount of O_2 through abiotic sources, generating a "false positive" biosignature. A review of several of these scenarios and the means by which they can be distinguished from biological sources can be found in Meadows et al. (2018b).

Generally speaking, abiotic oxygen is produced by the splitting of H_2O or CO_2 and the subsequent escape of hydrogen to space. Water vapor can be split either in the upper atmosphere due to UV-driven photolysis (e.g., Wordsworth and Pierrehumbert, 2014; Luger and Barnes, 2015; Meadows et al., 2018a; Wordsworth et al., 2018) or on the surface through a photocatalytic reaction involving titanium dioxide (Narita et al., 2015). The photolysis of carbon dioxide can lead to a buildup of abiotic O_2, which is exacerbated in the radiative environment of low-mass stars, or when the outgassing flux of reducing species (namely H_2 and CH_4) is much lower than on Earth (Hu et al., 2012; Domagal-Goldman et al., 2014; Tian et al., 2014; Gao et al., 2015; Harman et al., 2015; Hu et al., 2020).

More optimistically, other models show that the lightning-driven recombination of CO and O (Harman et al., 2018) as well as volcanic outgassing at rates comparable to Earth's (e.g., Hu et al., 2012; James and Hu, 2018) could each counter O_2 buildup from abiotic sources. Further research into such preventative factors may alleviate concerns about the reliability of this biosignature.

Some of these proposed abiotic sources of O_2 could, in principle, be present on both inhabited and non-inhabited planets, and would mask the age-oxygen correlation by imbuing planets with O_2 from a young age. To investigate this, we allow for some age-independent fraction of EECs to have abiotic oxygen sources, regardless of whether or not they have life. For most of the results presented below we set this parameter to zero

so that there are no planets with abiotic O_2, but we investigate their impact in Section 5.3.3.

5.2.3 Stellar age distribution

Recent studies have found that while the star formation history of the Milky Way disk has varied measurably, it is to first order uniform in age, with the oldest disk stars forming ~ 10 Gya (e.g., Snaith et al., 2015; Fantin et al., 2019; Mor et al., 2019). Assuming homogeneous star formation throughout the disk, the age distribution of nearby stars should resemble this history, excepting a small deficit of old, massive stars with main-sequence lifetimes shorter than 10 Gyr. We therefore assume a uniform stellar and planet age distribution from $0-10$ Gyr for most of our results.

However, it may be more efficient to prioritize observing only young and old stars, so as to maximize the difference in $f_{O_2}(t)$ for the models in Figure 5.1. This is especially important for surveys limited by available telescope time, where the target list must necessarily be pruned. Likewise, excluding young or old stars from any survey could inhibit the detectability of the age-oxygen correlation. To examine these considerations, we also simulate samples consisting of different combinations of young (0–2 Gyr), intermediate (2–7 Gyr) and old (7–10 Gyr) age ranges. We discuss the impact of the age distribution in Section 5.3.2.

5.2.4 Correlation test

To determine whether the age-oxygen correlation can be confidently identified in our simulated samples, we apply two statistical tests, each implemented using SciPy (Virtanen et al., 2020). The first is the Mann-Whitney U test (Mann and Whitney, 1947), which is applied to measurements of a variable (i.e., age) from two independent populations (i.e., oxygenated and anoxic planets), to determine whether the age distribution of one population is stochastically greater than the other. Unlike the similar and more widely used Student's t-test, the Mann-Whitney test does not assume the two age distributions to be of normal shape and equal variance, and compares the samples through their mean ranks

rather than through their sample means. We also apply Spearman's rank correlation test (e.g., Wall and Jenkins, 2003), which is usually applied to detect non-linear, monotonic correlations between two variables. The two variables can be continuous (i.e., age) or discrete (i.e., whether the planet has O_2).

We use the Mann-Whitney test to calculate most of our results, as we believe its underlying assumptions most accurately match our data, but we also compare its efficiency to that of Spearman's test in Section 5.3.4. Each of the tests reports a p-value representing the probability that age and oxygen are *not* apparently correlated, with values $p < 0.05$ corresponding to a significant (95% likely) correlation. Finally, since we are testing a directional hypothesis (that age and oxygen are *positively* correlated), we calculate one-tailed p-values.
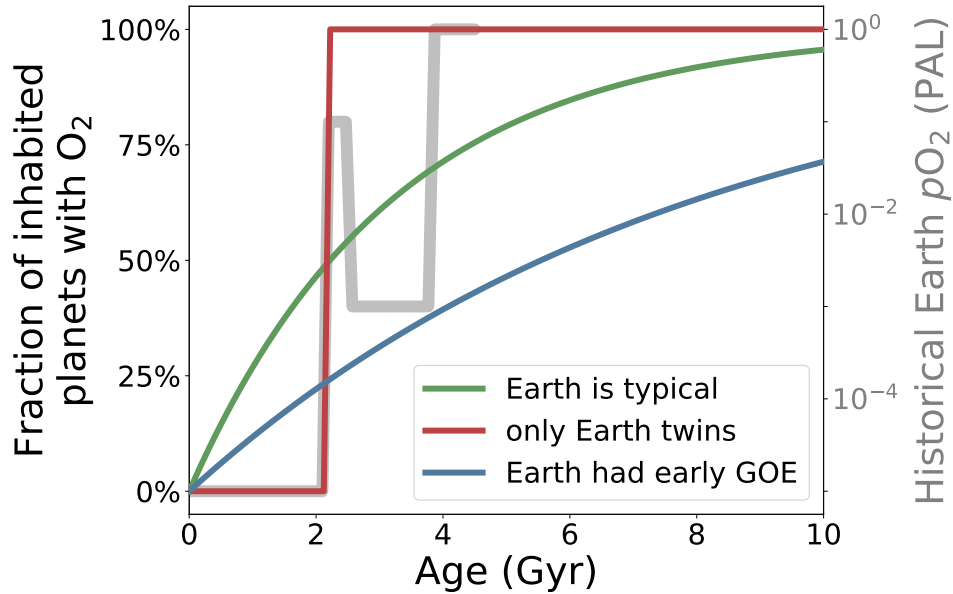
Figure 5.1: (In color, left axis) Three functions that we assume to describe $f_{O_2}(t)$, the fraction of inhabited planets that have a detectable amount of oxygen as a function of age t. In the first two cases (green and red), the typical inhabited planet undergoes a GOE at the same epoch as Earth's ($t_H \sim 2.2$ Gyr). (In gray, right axis) We plot an estimate of Earth's historical O_2 abundance versus the present atmospheric level (PAL), adapted from Reinhard et al. (2017). Note that estimates for pO_2 when Earth was younger than ~ 4 Gyr range by up to two orders of magnitude.
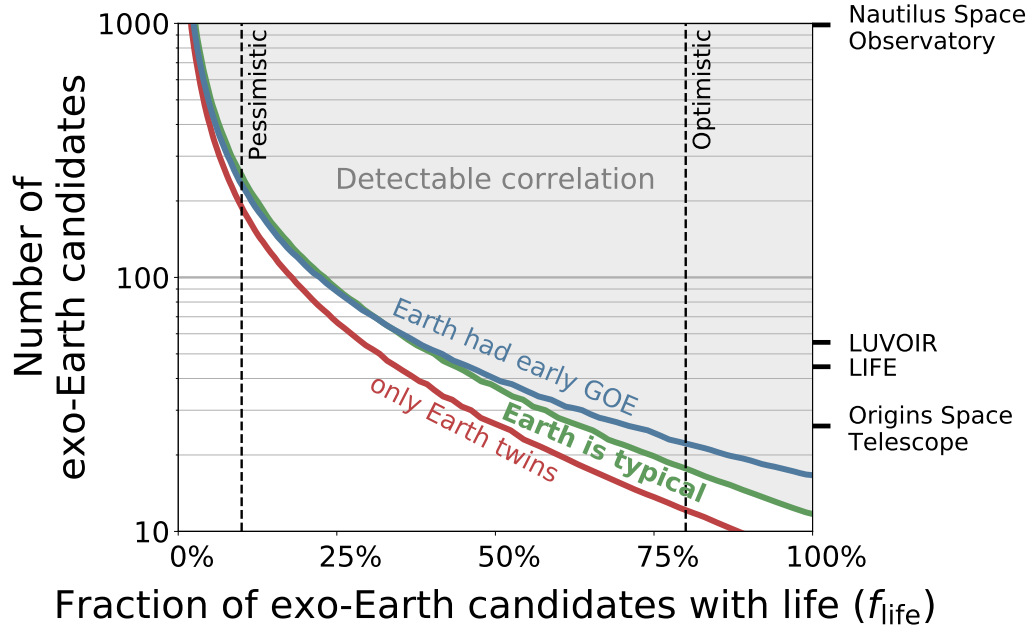
Figure 5.2: The number of EECs as a function of f_{life} that must be characterized to confidently ($p = 0.05$) detect a correlation between their ages and the fraction with O_2 or O_3 in their atmospheres. Results are plotted for each of the cases detailed in Figure 5.1 and a target sample with a uniform distribution of ages between $0 - 10$ Gyr. The dashed lines indicate "pessimistic" and "optimistic" cases for the frequency of inhabited worlds among EECs. On the right, we include estimates for the EEC detection yield of a few possible future observatories reviewed in Section 5.4.1. Note, however, that these estimates were calculated using different methods, and the occurrence rates used to determine them may have been overestimated (Pascucci et al., 2019).

5.3 Results

5.3.1 How many planets must be observed to detect a correlation?

The contour plots in Figure 5.2 delineate the number of planets - as a function of f_{life} - that must be characterized to detect the age-oxygen correlation with high confidence ($p < 0.05$). The shaded area marks the region of the parameter space in which the proposed age-oxygen correlation could be identified. Each coloured contour corresponds to one of the cases in Figure 5.1 – except where otherwise stated, we consider "Earth is typical" as our baseline case.
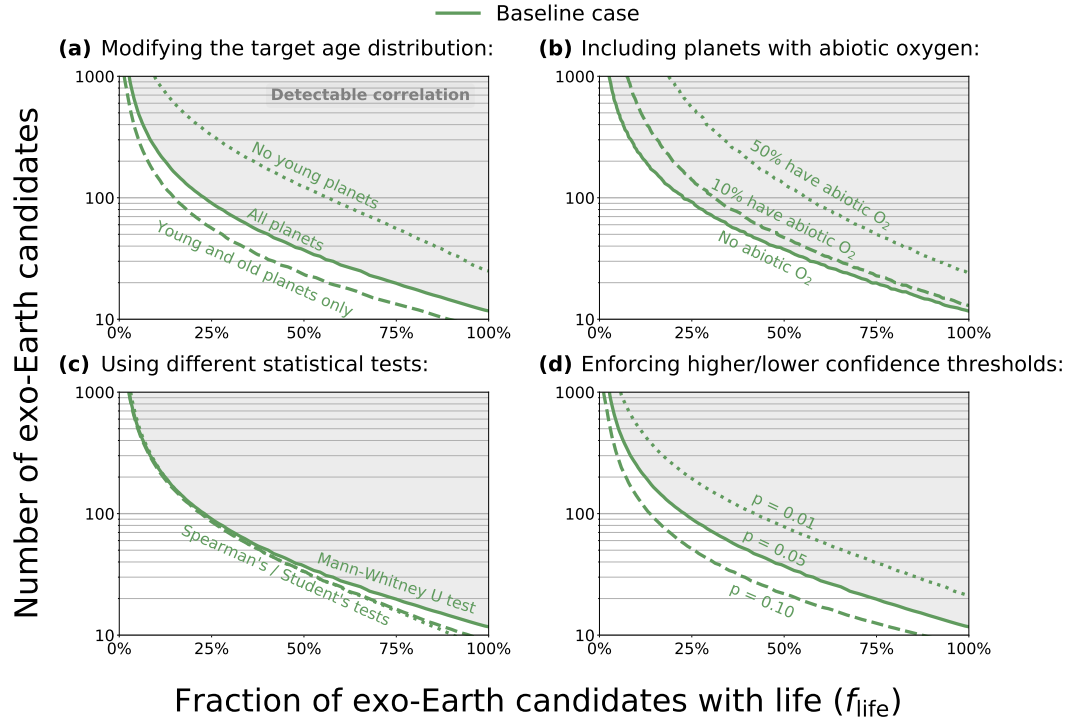
Figure 5.3: We re-calculate the number of EECs that must be characterized to detect the age-oxygen correlation in the "Earth is typical" case, but under different conditions than assumed in Figure 5.2. For comparison, we re-plot the baseline result in Figure 5.2 as a solid line in each panel. (a) We modify the age distribution of the target sample by including only young (0–2 Gyr) and old (7–10 Gyr) planets (dashed) or by excluding young planets (dotted). (b) We assume that 10% (dashed) or 50% (dotted) of EECs have some abiotically-produced O_2. This fraction includes inhabited and non-inhabited planets alike. (c) We re-calculate the results using Spearman's rank correlation test (dashed) and Student's t-test (dotted) to determine the detection significance. (d) We plot contours for a less ($p = 0.1$; dashed) or more ($p = 0.01$; dotted) confident detection of the age-oxygen correlation.

For the baseline case, we find that a sample of ~ 20 EECs would be sufficient to detect the age-oxygen correlation if life is present on 80% of such worlds. In a more pessimistic case where life exists on only 10% of these planets, then ~ 300 characterizations are required. If life is present on fewer than 2% of EECs, then the sample size required exceeds one thousand.

5.3.2   What is the optimal age distribution of target stars for this experiment?

To maximize the science yield of missions capable of detecting biosignatures, it may be prudent to prioritize their targets based on age. In Figure 5.3(a) we investigate the impact of different age-based target selection strategies on a survey's ability to test for an age-oxygen correlation.

We find that by selecting targets with ages just between 0–2 and 7–10 Gyr, the total number of planets required to detect the age-oxygen correlation drops by about 35%. Alternatively, when a survey excludes young systems, the number of planets required increases by a factor of ~ 3. These results suggest that, when possible, surveys of habitable exoplanets should not prioritize intermediate and older age planets over younger targets. In fact, the youngest planets are the most important targets for testing our hypothesis, and should be prioritized in surveys that are limited by observing time rather than fundamental instrument constraints.

5.3.3   What is the impact of abiotic sources for O_2?

For most of our results, we assume that all oxygen is produced by life, but in Figure 5.3(b) we investigate the amount of noise introduced if some age-independent fraction of EECs have abiotic O_2 sources. We find the impact on the required sample size to be modest to considerable. If abiotic O_2 exists on about 10% of EECs, then the required sample size must increase by 25-100% depending on f_{life}. In the pessimistic case that half of EECs have abiotic O_2 and life is rare, the required sample size increases by an order of magnitude.

5.3.4 What is the most efficient test for detecting the age-oxygen correlation?

In Figure 5.3(c) we compare results using each of the statistical tests reviewed in Section 5.2.4, as well as the more commonly-used Student's t-test (which is not strictly applicable to our case as it assumes normal age distributions). Generally speaking, the three tests yield similar results. While Student's and Spearman's tests are slightly more sensitive to the correlation for high values of f_{life}, the assumptions of the Mann-Whitney test most accurately match our data set, which consists of two non-normal age distributions (for oxygenated and anoxic atmospheres) with unequal variances.

While we have chosen $p = 0.05$ (i.e., a 5% probability that age and oxygen are uncorrelated) as our threshold for a confident detection of a positive correlation, a lower or higher confidence level detection may be achieved through characterizing fewer or more EECs. In Figure 5.3(d) we calculate the number of targets required to detect the age-oxygen correlation with low confidence ($p < 0.1$) or very high confidence ($p < 0.01$). These contours demonstrate that while a confident detection may be out of reach for a given sample size if life is rare, preliminary evidence can still be acquired to motivate a more in-depth survey.

5.4 Discussion

5.4.1 Future observatories could test the proposed age-oxygen correlation

Current observatories lack the capability to detect oxygen or ozone absorption in the atmospheres of terrestrial exoplanets. However, multiple ambitious space mission concepts that have been proposed in the literature could perform this characterization for statistically meaningful numbers of planets. Here, we identify several such concepts and compare their likely sample sizes (where available) to the requirements we predict in Figure 5.2.

Two observatories would use coronagraphic instruments to directly image potential exo-Earths orbiting nearby FGK stars. The Large UV/Optical/IR Surveyor (LUVOIR, The LUVOIR Team, 2019) would feature an 8- to 15-meter diameter segmented primary mirror, while the Habitable Exoplanet Observatory (HabEx, Gaudi et al., 2020) would

make use of a 4-meter monolithic mirror. HabEx would also launch with a starshade, which would maneuever separately from the telescope to occult the targeted host star and enable deep spectroscopic characterization of its planets. With their broad UV-to-NIR wavelength coverage, both telescopes could probe for ozone absorption at far-UV wavelengths as well as O_2 in the visible spectrum. Employing the yield optimization methodology first developed by Stark et al. (2014), both concept studies have reported estimates for the number of EECs that they could detect, namely 54^{+61}_{-34} for LUVOIR and 8^{+9}_{-5} for HabEx (see also Stark et al., 2015; Stark et al., 2016; Stark et al., 2019; Kopparapu et al., 2018). As such, LUVOIR would be able to test for the age-oxygen correlation given $f_{\text{life}} > 50\%$. HabEx would likely lack the sample size required to detect it on its own, but could contribute substantially towards building up a sufficiently large sample of characterized EECs if complemented by ground- or space-based efforts targeting nearby M dwarfs.

Three mission concepts would use transit and/or phase curve spectroscopy to characterize potential exo-Earths. The *Nautilus* Space Observatory (Apai et al., 2019) would consist of thirty-five unit space telescopes, each with an ~ 8.5-meter diameter ultralight, diffractive-refractive lens as the primary light-collecting element. Through visible-to-NIR transit spectroscopy, the array could be used to search for oxygen or ozone in the atmospheres of up to one thousand EECs. With such a sample, *Nautilus* could test the age-oxygen correlation even in the pessimistic case where only 10% of EECs are inhabited. Two other telescopes would enable mid-infrared transit and phase curve spectroscopy of planets orbiting mid-to-late M dwarfs, including the Origins Space Telescope (~ 26 characterized EECs, Meixner et al., 2019) and the Mid-Infrared Exoplanet Climate Explorer (Staguhn et al., 2019), and would infer the presence of oxygen through O_3 absorption between 9-10 μm. In the optimistic case, a sample of $20 - 30$ planets could achieve the requirements outlined in Figure 5.2, or it could be combined with the yield of missions targeting FGK stars.

Space-based infrared interferometry offers another avenue towards directly imaging exo-Earths through their thermal emission. The Large Interferometer For Exoplanets

(LIFE) project[3] aims to coherently combine light from four ~ 2.8-meter mirrors in order to detect and characterize nearby exo-Earths in the 5–25 μm wavelength range. LIFE would characterize the atmospheres of ~ 45 EECs orbiting nearby GKM stars, and with such a sample could detect the age-oxygen correlation for $f_{\text{life}} > 25\%$ (Kammerer and Quanz, 2018; Quanz et al., 2018).

A note of caution: the yield estimates above have been adapted directly from literature references, and may not be directly comparable as they were calculated using different techniques. All of them rely on estimates of $\eta_\oplus$ (the number of potentially habitable planets per star) extrapolated from *Kepler* data, but recent work by Pascucci et al. (2019) suggests that such estimates are exaggerated by a factor of 4–8$\times$, as the *Kepler* radius distribution of short-period planets is heavily impacted by atmospheric loss. Nevertheless, the numbers cited suggest that testing for the age-oxygen correlation may be an achievable science goal for some of these missions.

In summary, we conclude that future space missions that are being designed to detect oxygen in the atmospheres of individual planets could use the same capabilities to test the null hypothesis that the Earth's atmospheric evolution is typical for an inhabited world. To achieve this science goal, the stellar age distribution of the target sample should be carefully considered: when it is necessary to prioritize targets, surveys are encouraged to favor young stars to maximize their sensitivity to changes in atmospheric oxygen content during the first few billion years.

5.4.2   Impact of planet and stellar properties

In our analysis we assume that the amount of time required for a planet to reach a GOE-like transition is random, but typically between $1-10$ Gyr. In reality, this timescale will be dependent on each planet's properties. In order for a GOE-like event to occur, a planet's biological oxygen source must grow large enough and/or its oxygen sinks must become minimal enough that the former overwhelms the latter. Here we address ways in which these sources and sinks might vary across different types of habitable worlds.

---

[3]https://www.life-space-mission.com/

Catling et al. (2005) coin the term "oxygenation time" to refer to the time required for a planet to acquire enough atmospheric O_2 to support complex life, and they argue that a planet's size, composition, and the presence or absence of continents could affect this timescale in diverse ways. For example, larger planets, or planets with more reducing initial compositions, will have a greater inventory of reducing matter to exhaust. The interior heat flux could be higher on planets larger than Earth, or for tidally-heated planets on close-in orbits around low-mass stars (Driscoll and Barnes, 2015), which would affect outgassing rates. Continents play a role in both removing oxygen from the atmosphere (through outgassing and rock weathering) and replenishing it (through the burial of organic matter), but on a planet without continents these processes would be diminished (Lunine, 2013).

The escape of hydrogen from Earth's atmosphere into space acts to oxidize the planet, an effect that may have triggered the GOE since hydrogen escape from early Earth's atmosphere was likely much faster than at present (Catling et al., 2001; Claire et al., 2006; Zahnle et al., 2013, 2019). On Earth, the hydrogen escape rate is limited by its diffusion rate into the exosphere (Hunten and Donahue, 1976). The diffusion length decreases with surface gravity, which ranges from $\sim 0.3 - 2.5\ g$ on rocky exoplanets (e.g., Neil and Rogers, 2020), so diffusion-limited escape should be more efficient on larger planets - but at the same time the maximum escape rate of hydrogen from the exosphere will be throttled due to increased gravitational potential (Catling et al., 2005). In total, we might expect a non-monotonic relationship between hydrogen escape and planet size that could accelerate or inhibit the oxygenation of the atmospheres of other inhabited worlds.

Since the advent of oxygenic photosynthesis is presumed a prerequisite for global oxygenation, the rate at which evolutionary changes occur in the biosphere could also limit the oxygenation timescale. Several authors have investigated factors affecting the pace of biological evolution on planets hosted by low-mass stars, with some suggesting that prebiotic chemistry could be inhibited by a deficit of ultraviolet radiation (Buccino et al., 2007; Ranjan et al., 2017; Rimmer et al., 2018), and others proposing that complex life could take longer to subsequently evolve (Haqq-Misra and Kopparapu, 2018; Haqq-Misra, 2019). Biogenic oxygen levels may be substantially lower around low-mass stars,

as they emit less of the visible-wavelength radiation that drives oxygenic photosynthesis on Earth (e.g., Kiang et al., 2007; Gale and Wandel, 2016; Lehmer et al., 2018; Mullan and Bais, 2018; Ritchie et al., 2018; Lingam and Loeb, 2018, 2019). The net effect of these factors could be to substantially delay or entirely prevent the oxygenation of the atmosphere for habitable planets orbiting low-mass stars.

A planet's size, bulk composition, and its host star's spectral type can often be constrained through observation, so it is conceivable that one could control for their effects - but this would necessitate a larger sample. It may be possible to detect continents by measuring a planet's photometric variability through extensive direct imaging observations (e.g. Ford et al., 2001; Cowan et al., 2009; Cowan and Fujii, 2018; Farr et al., 2018; Lustig-Yaeger et al., 2018; Fan et al., 2019; Aizawa et al., 2020), but likely only for a limited number of optimal targets.

In total, we expect that the diversity of planet compositions and environments will result in a corresponding diversity of oxygenation timescales, but overall the age-oxygen correlation should remain as long as some planets were oxygenated within their first several Gyr. Our analysis accommodates this diversity by simulating planets with such timescales ranging from $\sim 1 - 10$ Gyr, but if the typical timescale is in fact very short (< 1 Gyr) or long (> 10 Gyr), then a larger sample size will likely be required to detect the correlation. Further detailed theoretical treatments of the variation of oxygen sources and sinks across a realistic range of planetary properties will be valuable for evaluating the assumptions made in Figure 5.1.

5.4.3 Verifying O_2 as a potential biosignature

We propose that a potential future discovery of a positive age-oxygen correlation would serve as additional evidence for life on oxygen-bearing planets, even if concerns about false positives cannot be ruled out through contextual evidence for individual planets in the sample (e.g., Meadows et al., 2018b). While multiple mechanisms have been proposed for the abiotic generation of O_2, none so far have been shown to produce a positive correlation of O_2 content with age. On the other hand, such a correlation does reflect the history of biogenic oxygen in Earth's atmosphere, so if discovered it would suggest a

similar (i.e. biological) history for other worlds. Nevertheless, like all proposed biosignatures, the discovery of an age-oxygen correlation would need to be rigorously scrutinized to ensure that no plausible abiotic evolutionary scenarios could produce it.

On the other hand, if enough planets with oxygenated atmospheres are detected that a positive age-oxygen correlation can be *ruled out* over Gyr timescales, this would imply that most planets that already have oxygen acquired it before ~ 1 Gyr. An explanation would be required for why Earth took substantially longer to become oxygenated than most oxygen-rich planets. Such an explanation may be found through modeling the effects of planet and stellar properties on redox balance as discussed in Section 5.4.2, or perhaps through an argument about anthropic bias (i.e., if complex and intelligent life is more likely to evolve on planets with late GOEs, then our planet is more likely to have had a late GOE).

5.4.4 Complicating factors for detecting O_2 or O_3

We assume that O_2 or at least O_3 will be detectable for every post-GOE planet, but complicating factors could make this assumption optimistic. Clouds and hazes can mask absorption by low-altitude gasses, an issue that is expected to affect both direct imaging and transit observations of rocky planets (e.g., Arney et al., 2017; Rugheimer and Kaltenegger, 2018; Wang et al., 2018; Kawashima and Rugheimer, 2019; Lustig-Yaeger et al., 2019b). For transit spectroscopy, it could prove difficult to disentangle the spectral features of the atmosphere from those of the stellar photosphere (Apai et al., 2018b). This may inhibit the detection of biosignatures on planets around M dwarfs (Rackham et al., 2018; Zhang et al., 2018; Iyer and Line, 2020), but would be a less prominent issue for FGK stars (Rackham et al., 2019b). Detailed spectral modeling of the photosphere may also help to resolve this degeneracy (Pinhas et al., 2018; Rackham et al., 2019a; Wakeford et al., 2019; Iyer and Line, 2020).

We again emphasize that even in cases where O_2 is difficult to observe (e.g., due to clouds), strong O_3 absorption may still be visible. Generally, while these complicating factors may have the effect of increasing the sample size required to detect the age-oxygen correlation, they in principle should not inhibit it.

5.4.5 Assumed correlations

In our analysis we consider three functions to describe the fraction of inhabited planets with oxygenated atmospheres. The first and third functions have the following exponential form:

$$f_{O_2}(t) = 1 - \exp(-t/\tau)$$

This is appropriate if we assume that every inhabited planet has an equally small probability to undergo a GOE during each consecutive interval of time during its history. If Earth is a typical example, then this probability is $\sim 3\%$ per 100 Myr. This case seems appropriate if the timing of the GOE is set by changes in the biosphere, as several random and independent evolutionary steps must occur before oxygenic photosynthesis can become a dominant form of metabolism.

In the second case we assume a step function:

$$f_{O_2}(t) = \begin{cases} 1 & t \geq 2.2 \text{ Gyr} \\ 0 & t < 2.2 \text{ Gyr} \end{cases} \tag{5.1}$$

Under this case, every planet is an exact Earth analog and undergoes a GOE at the same age as did Earth. While this is unlikely to actually be correct, a step function would be the most easily detectable correlation, so our results for this case reflect a lower limit on the required sample sizes.

Despite the difference between the functional forms of the first/third and second cases, the sample sizes that must be observed to test them typically agree to within a factor of two, suggesting that our results are relatively consistent across these cases.

5.4.6 Prospects for determining stellar ages

For simplicity, we assume that each planet's age (which is approximately the age of its host star) is known with high precision, but this is not a true assumption for most known exoplanets today.

It is plausible that precise (< 1 Gyr) age constraints could be achieved through asteroseismology, as demonstrated for several *Kepler*/K2 planet hosts (e.g., Mathur et al.,

2012; Chaplin et al., 2014; Silva Aguirre et al., 2015; Creevey et al., 2017; Kayhan et al., 2019; Lund et al., 2019). With the same method, the PLATO mission could allow for 10% precision age measurements of hundreds of bright solar-type stars (Rauer et al., 2014).

However, asteroseismic pulsations have so far proved difficult to detect in low-mass stars despite extensive efforts (Baran et al., 2011b,a; Krzesinski et al., 2012; Baran et al., 2013; Rodríguez-López et al., 2015; Rodríguez et al., 2016; Berdiñas et al., 2017). For now, age estimates for low-mass stars rely on various spectroscopic and photometric relations. For example, Burgasser and Mamajek (2017) combine several diagnostics to determine the age of TRAPPIST-1, a well-studied ultra-cool dwarf known to host multiple potentially habitable planets (Gillon et al., 2017). Despite a thorough analysis, the authors are only able to constrain the age of the system with a 1σ precision of ± 2.2 Gyr, demonstrating that securing sub-Gyr age constraints for low-mass stars is not feasible using existing techniques.

Provided systematic errors are minimal, age uncertainties could be factored into a statistical correlation test. Doing so, however, would likely increase the required sample size if the uncertainty is much larger than the age range over which the correlation is expected ($\gtrsim 1$ Gyr). It is therefore important that advances be made over the next two decades in the measurement of stellar ages, particularly for low-mass stars.

5.4.7 Luminosity evolution for low-mass stars

A potential source of uncertainty comes from the luminosity evolution of the host star, and therefore the evolution of its habitable zone. Habitable zone planets around low-mass stars are the best targets for characterization through transit spectroscopy because they produce larger relative transit depths, are more likely to transit, and transit more often than Earth twins around Sun-like stars. However, their host stars' habitable zones contract significantly during the pre-main sequence phase, which can last for hundreds of Myr. Planets towards the inner edge may have regained their habitability only after the star reached the main sequence, or may have lost it permanently (e.g. Ramirez and Kaltenegger, 2014; Tian and Ida, 2015; Luger and Barnes, 2015; Barnes et al., 2016).

To account for this caveat, more detailed future studies could estimate and subtract the

amount of time for which the planet was outside of the habitable zone from its age before performing a correlation test. In contrast, planets that have been rendered permanently uninhabitable by the pre-main sequence star represent a reduction in f_{life}, and are therefore already factored into our analysis.

5.5 Conclusions

Motivated by a new generation of space missions concepts that aim to search for atmospheric biosignatures across statistically meaningful samples of planets, we explore how constraints on the presence of atmospheric O_2 or O_3 as a function of age could be used to study how inhabited planets and life co-evolve, and to test the robustness of oxygen as a biosignature. A possible starting hypothesis for the evolution of inhabited planets is that their atmospheres evolve in a similar manner to Earth's. We show that this null hypothesis predicts a strong, positive age-oxygen correlation among such worlds. The presence of such a trend, if detected by future observatories, could serve as additional evidence for life.

We show that by detecting or rejecting the presence of O_2 – or its byproduct O_3 – in a sufficiently large sample of potentially habitable planets, it will be possible to confirm this hypothesis using statistical correlation tests – without needing to know which planets are inhabited. To confidently detect the age-oxygen correlation, we find that a sample size of ~ 20 potentially habitable planets must be observed if $\sim 80\%$ of them are in fact inhabited. If the inhabited fraction is only $\sim 10\%$, then ~ 300 planets must be observed, and if it is smaller than $\sim 2\%$ then more than one thousand planets are required. These sample sizes are similar to those predicted for ambitious space missions proposed to launch within the coming decades.

Our results have important implications for the target selection of future biosignature surveys that should be considered as their missions are designed and built. Namely, surveys that must down-select or prioritize their target lists should favor young and old over intermediate-age stars, as doing so could reduce the number of planets required to detect the age-oxygen correlation by about 35%. Similarly, surveys should avoid excluding

young stars from their target lists, as this could increase the number of planets required by a factor of ~ 3. Such target selection strategies will require precise stellar age measurements, which should be achievable for many FGK dwarfs through asteroseismology, but are yet out of reach for low-mass stars.

These results are also sensitive to the abundance of abiotically-produced O_2 in exoplanet atmospheres. While Earth has no substantial non-biological oxygen sources, it has been proposed that a small fraction of potentially habitable exoplanets could. If this fraction exceeds $\sim 10\%$, then many more planets must be characterized to detect the age-oxygen correlation.

Our study offers a promising initial analysis of the capacity of next-generation observatories to study the oxygen evolution of habitable planets. Future studies can expand upon this groundwork by incorporating quantitative treatments of the influence of planet size, composition, stellar environment, and other factors that may impact the proposed age-oxygen correlation. Finally, while in-depth studies of individual planets – especially those presenting biosignatures – will be invaluable, statistical analyses could enable a broader understanding of life as a universal phenomenon. We encourage that this and other statistical hypotheses be given proper consideration as new telescopes and instruments are being designed to characterize the atmospheres of habitable exoplanets.

5.5.1 Acknowledgements

CHAPTER 6

Bioverse: A Simulation Framework to Assess the Statistical Power of Future
Biosignature Surveys

*This chapter has been published as Bixel and Apai (2021) and is reproduced here with
permission.*

Abstract

Next-generation space observatories will conduct the first systematic surveys of terrestrial exoplanet atmospheres and search for evidence of life beyond Earth. While in-depth observations of the nearest habitable worlds may yield enticing results, there are fundamental questions about planetary habitability and evolution that can only be answered through population-level studies of dozens to hundreds of terrestrial planets. To determine the requirements for next-generation observatories to address these questions, we have developed `Bioverse`. `Bioverse` combines existing knowledge of exoplanet statistics with a survey simulation and hypothesis testing framework to determine whether proposed space-based direct imaging and transit spectroscopy surveys will be capable of detecting various hypothetical statistical relationships between the properties of terrestrial exoplanets. Following a description of the code, we apply `Bioverse` to determine whether an ambitious direct imaging or transit survey would be able to determine the extent of the circumstellar habitable zone and study the evolution of Earth-like planets. Given recent evidence that Earth-sized habitable zone planets are likely much rarer than previously believed (Pascucci et al., 2019), we find that space missions with large search volumes will be necessary to study the population of terrestrial and habitable worlds. Moving forward, `Bioverse` provides a methodology for performing trade studies of future observatory concepts to maximize their ability to address population-level questions, including and beyond the specific examples explored here.

6.1 Introduction

The field of exoplanet science stands at an exciting turning point. In the past, most exoplanet surveys aimed only to constrain bulk properties - such as size, period, and mass. Moving forward, several groups are developing concepts for space telescopes that would enable the atmospheric characterization of temperate terrestrial planets. Such concepts include the Large UV/Optical/Infrared Surveyor (LUVOIR, The LUVOIR Team, 2019), the Habitable Exoplanet Observatory (HabEx, Gaudi et al., 2020), the Origins Space Telescope (Meixner et al., 2019), the Nautilus Space Observatory (Apai et al., 2019), the Large Interferometer for Exoplanets (LIFE, Quanz et al., 2018), and the Mid-Infrared Exoplanet Climate Explorer (MIRECLE, Staguhn et al., 2019). By looking for biosignatures in the atmospheres of temperate Earth-sized planets, these observatories would conduct the first systematic search for life beyond the Solar System.

Next-generation observatories will be able to study some of the closest terrestrial exoplanets in unprecedented detail, but this is only the start of their scientific capability: observatories that can study tens to hundreds of terrestrial planets will allow for the first statistical constraints on the atmospheric, geological, and biological properties of terrestrial planets. Some recent works have explored statistical trends and patterns that may only be evident at the population level. For example, habitable zone models predict patterns in atmospheric CO_2 and H_2O abundance (Bean et al., 2017; Lehmer et al., 2020) as well as color and albedo across a range of stellar insolations (Checlair et al., 2019). Venus analogs may have larger apparent radii than their temperate siblings due to their thick, post-runaway greenhouse atmospheres (Turbet et al., 2019). Earth's geological record suggests a possible relationship between the ages and oxygen content of Earth-like planets, assuming their atmospheres evolve similarly to Earth's (Bixel and Apai, 2020b), and with a large enough sample size of potentially habitable planets, next-generation surveys could place the first constraints on the frequency of life in the universe (Checlair et al., 2021). An understanding of population-level trends will provide context for the interpretation of possible biosignatures on individual worlds and could illuminate their potential false positive (i.e. non-biological) sources (Apai et al., 2017; Meadows et al., 2018b). To

avoid statistical false positive scenarios, efforts must also be made to understand which distinct mechanisms could produce the same apparent trends. For example, an increase in cloud deck altitude with insolation could masquerade as a signature of atmospheric erosion in a sample of transiting exoplanets (Lustig-Yaeger et al., 2019a).

Recent research has identified key outstanding questions about terrestrial exoplanets, their planetary systems, and the processes that shape them for which future observatories might provide insights (see the SAG 15 report for an overview of several such questions in the context of direct imaging missions, Apai et al., 2017). For example: what are the processes that shape their atmospheric loss (e.g., Zahnle and Catling, 2017)? Is the habitable zone wide (e.g., Kasting et al., 1993; Kopparapu et al., 2014) or narrow (e.g., Hart, 1979)? What is the relationship between planet size and tectonic activity (e.g., Valencia et al., 2007; Dorn et al., 2018)? Are habitable planets equally common around stars of different mass and activity levels (e.g., Shields et al., 2016)? Which, if any, of these questions could be answered with a next-generation observatory will depend on its technical design and observing strategy. One important metric is the number of terrestrial habitable zone planets that it could realistically detect, but only a subset of these will be habitable, and even inhabited worlds may vary substantially from Earth in their atmospheric composition and evolutionary history. Furthermore, deep spectroscopic characterization of individual planets will be time-consuming, so strategic choices must be made as to which planets to characterize and at what wavelengths. For these reasons, analyses based solely on the detection yield predictions of future space mission concepts will provide an optimistic assessment of their statistical power.

To enable meaningful statistical hypotheses that can be tested by future observatories, we have developed `Bioverse`. `Bioverse` estimates the statistical power of next-generation exoplanet surveys to detect and study population-level trends by simulating the underlying planet population, survey limitations, observing biases, and statistical analyses that a future observer would perform on a large set of observations of terrestrial planets. After the following brief description of the code structure, we describe its three main components in Sections 6.3 through 6.5. In Sections 6.6 and 6.7, we use `Bioverse` to determine the requirements for next-generation surveys to test the habitable zone concept

and study the evolution of Earth-like planets.

6.2 Code outline

`Bioverse` consists of three components, outlined in Figure 6.1. The first component generates planetary systems with bulk properties (e.g., size and period) drawn from statistical distributions, then applies theoretical models or parametric relationships to generate secondary properties of interest (e.g., atmospheric composition). The second component is a survey simulator that conducts observations of the simulated exoplanet population in direct imaging or transit spectroscopy mode. The survey simulator first determines which planets could be characterized within a finite allotted observing time, then generates a simulated data set representative of the telescope and instrument capabilities. The third component is a Bayesian framework that uses simulated datasets to test statistical hypotheses and estimate model parameters. By iterating through these components, we can use `Bioverse` to determine the statistical power of a proposed observatory to test different hypotheses.

`Bioverse` is written in Python[1] and designed for flexibility, so that different statistical assumptions and testable hypotheses can be implemented in the future. The specific set of assumptions that `Bioverse` is currently based on are listed in Table 6.1. Given the large number of parameters involved in `Bioverse`, we provide a table of abbreviations and symbols used in the text in Appendix B.1.

---

[1]A current version of the code can be found on GitHub, while the version used in this paper is archived on Zenodo (Bixel and Apai, 2021).
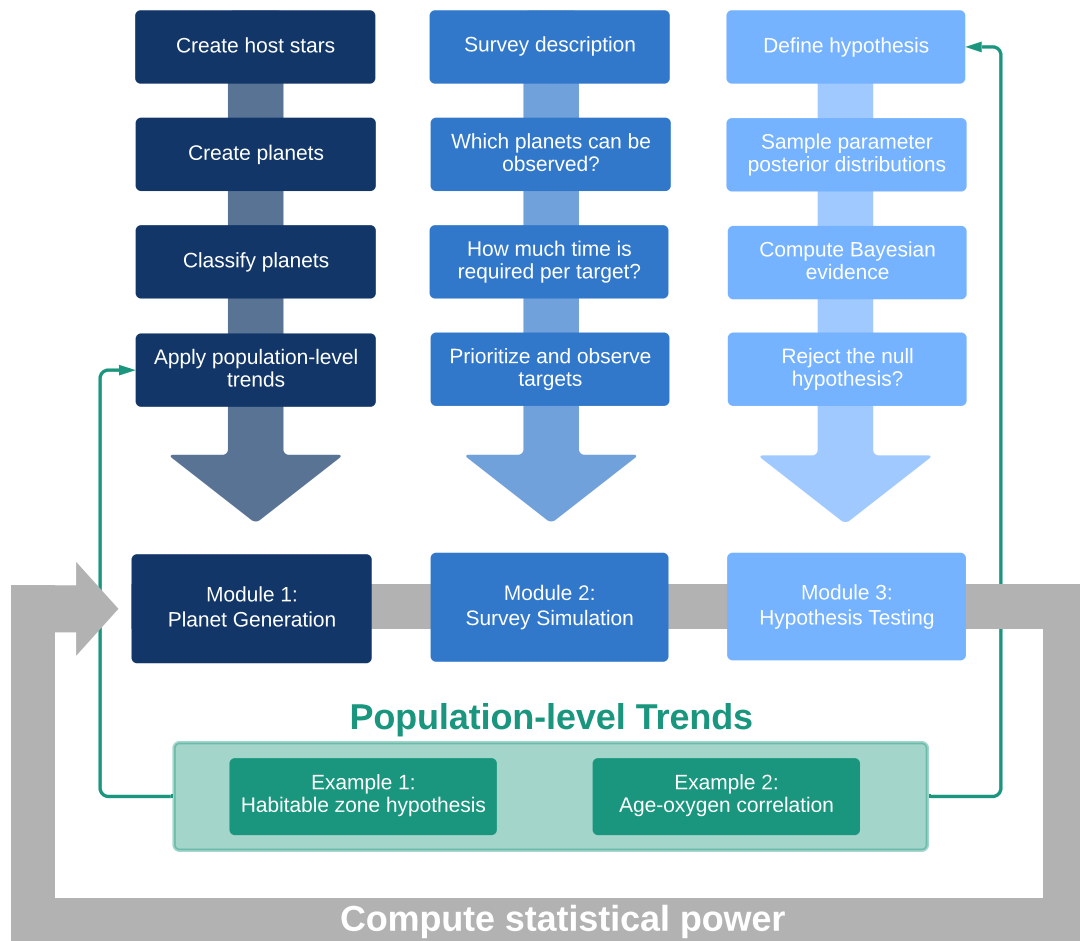
Figure 6.1: A high-level outline of the `Bioverse` code. In this paper, we apply `Bioverse` to assess the detectability of two hypothetical population-level trends (green) with next-generation survey telescopes. These relationships are injected into the simulated planet population by the first module, then tested as statistical hypotheses by the third module.

Table 6.1: Summary of Statistical Assumptions and Modeling Choices in `Bioverse`, with Associated References.

Topic	Assumptions	References
Host star distribution and properties	(Imaging mode) LUVOIR-A optimized target catalog	The LUVOIR Team (2019) and C. Stark (private correspondence)
	(Transit mode) Stellar mass function	Chabrier (2003)
	Main sequence mass-radius-luminosity relations	Pecaut and Mamajek (2013)
Planet occurrence rates	SAG13 occurrence rates, with modifications:	
	- $\eta_\oplus \approx 7.5\%$ for G stars (down from $\approx 24\%$)	Pascucci et al. (2019); Neil and Rogers (2020)
	- More planets around lower-mass stars	Mulders et al. (2015a,b)
Exo-Earth candidates	approximately Earth-sized ($0.8S^{0.25} < R < 1.4R_\oplus$) within the circumstellar habitable zone	various (see Section 6.3.4) K14
Observatory templates	(Imaging mode) 15-meter LUVOIR-A observatory	The LUVOIR Team (2019)
	(Transit mode) 50-meter equivalent Nautilus Space Observatory	Apai et al. (2019a)

Table 6.1: (continued)

Topic	Assumptions	References
Target prioritization	Finite observing time with overheads Observe in order of required time Prioritize targets to reduce survey biases	
Measurement noise	Photon-noise limited observations with characteristic wavelength λ_{eff} Required exposure time scales with distance, stellar brightness, and signal strength	
Model comparison	Compare alternative to null hypothesis through Bayesian evidence $\mathcal{Z}$ Significant evidence when $\Delta(\mathcal{Z}) > 3$ (where applicable) Frequentist comparison tests (e.g., t-tests)	

6.3 Planet generation

The first component of `Bioverse` creates simulated planetary systems around host stars in the solar neighborhood with a period and radius distribution informed by *Kepler* statistics. Other planet properties (such as mass and geometric albedo) are derived from empirical relationships or best-guess prior distributions. Finally, the simulated planet properties reflect the effects of hypothetical population-level trends that could be uncovered by a future survey of terrestrial planets.

6.3.1 Stellar properties

We begin by considering which stars in the solar neighborhood would be targeted by future biosignature surveys. Our strategy for simulating stellar systems is mass-dependent, and therefore depends on the observing technique used by the simulated survey. `Bioverse` currently considers observations through coronagraphic direct imaging (in "imaging mode") and transit spectroscopy (in "transit mode").

Direct imaging surveys will primarily target the habitable zones of higher-mass (FGK) stars within the nearest 30 pc, the majority of which have already been cataloged by space-based astrometry missions. Not all of these will be equally valid targets, due to the combined effects of distance and background noise sources, such as zodiacal dust (Stark et al., 2019). Sophisticated simulations for the LUVOIR mission concept (The LUVOIR Team, 2019) have produced an optimized list of targets whose habitable zones could feasibly be probed for Earth-like planets. In imaging mode, we use an optimized stellar target list for the 15-meter LUVOIR-A concept as the basis for simulating nearby planetary systems (C. Stark, private correspondence).

A survey of transiting habitable zone planets would be most sensitive to planets around low-mass (K and M) stars, as their habitable zone planets are more likely to transit, transit more frequently, and produce a deeper relative transit depth. However, the census of low-mass stars is not complete out to ~ 100 pc. Therefore, in transit mode, all stellar masses are randomly drawn from a present-day stellar mass function (Chabrier, 2003) and distribute them isotropically in space. We do not include any known stars or transiting

planets in the transit mode sample; as most nearby transiting planets remain undiscovered, this would have little effect on the overall statistical distribution of host star properties.

In both imaging and transit modes, we relate the stellar mass (M_*) to its radius, luminosity, and effective temperature (R_*, L_*, T_*) by interpolating a list of these properties as a function of spectral type (Pecaut and Mamajek, 2013). Each star is assigned an age drawn uniformly from 0–10 Gyr, reflecting the (to first order) constant star formation rate in the Milky Way for the past 10 Gyr (e.g., Snaith et al., 2015; Fantin et al., 2019; Mor et al., 2019).

6.3.2 Period and radius occurrence rates

Kepler has provided excellent insights into the frequency of planets as a function of period and size for a wide range of host stars. However, these statistics are only complete to periods $\lesssim 100$ days, and as such do not reach the habitable zone of Sun-like stars. As a result, estimates of $\eta_\oplus$ (the average number of habitable zone Earth-sized planets per star) have so far been based on extrapolation and are therefore model-dependent. NASA's Exoplanet Program Analysis Group chartered Science Analysis Group 13 (hereafter SAG13) to consolidate the results of several studies of *Kepler* occurrence rates into a single set of estimates for community use,[2] resulting in an oft-cited value of $\eta_\oplus \approx 24\%$ for G stars. Here, and elsewhere in this paper, the value of $\eta_\oplus$ uses the habitable zone model of Kopparapu et al. (2014) (hereafter K14; 0.95 − 1.67 AU for an Earth twin). We use the SAG-13 consensus occurrence rate power laws as the basis for determining the number, radii (R_p), periods (P), semi-major axes (a), and insolations (S) of planets in each system. However, the SAG13 metastudy was based largely on studies published before 2017, many of which did not assess planet occurrence as a function of stellar mass. We make the following two modifications to the SAG13 rates to reflect recent work.

First, we unilaterally decrease the number of planets per star by a factor of 3.2, such that $\eta_\oplus \approx 7.5\%$ for G stars. This is in response to the findings of Pascucci et al. (2019) that Earth-sized planets are more common at shorter orbital periods ($P \lesssim 25$ d) than in the habitable zone, which they ascribe to the effects of photoevaporation. Specifically,

---

[2]see this URL as well as Kopparapu et al. (2018)
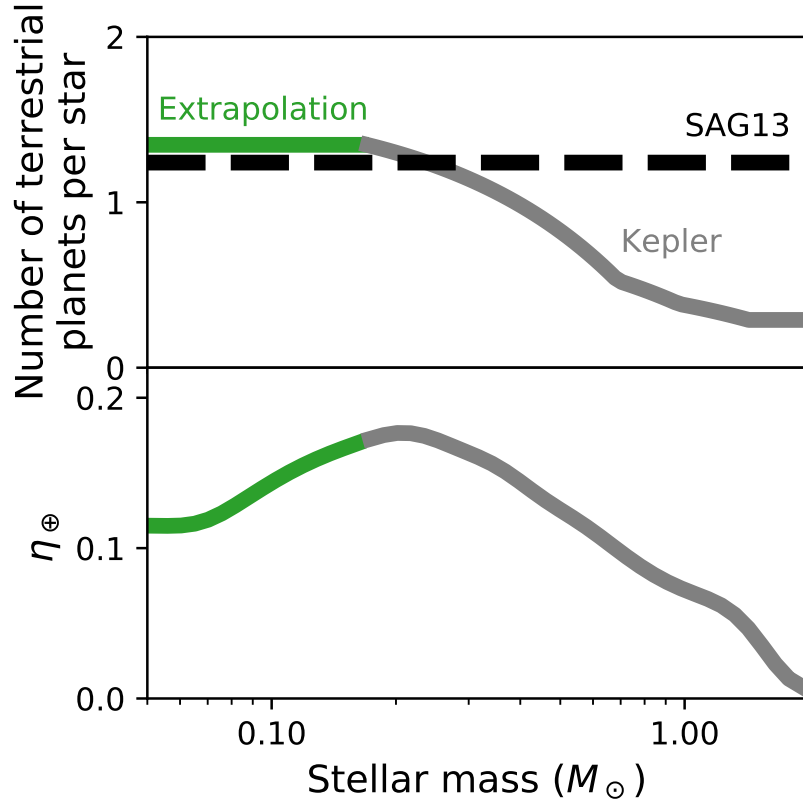
Figure 6.2: (Top) The assumed number of approximately Earth-sized planets ($0.7 < R < 1.5\,R_\oplus$) with orbital periods shorter than 3 yr per star, as a function of stellar mass. We modify the SAG13 estimate (black) by decreasing the overall planet count by $\sim 3\times$ and increasing the number of planets orbiting *Kepler* low-mass stars, as well as shortening their orbital periods (gray). We conservatively assume the occurrence rates to plateau for ultra-cool dwarfs (green). (Bottom) The corresponding value of $\eta_\oplus$ using the habitable zone model of K14.

they argue that a large fraction of Earth-sized planets on close-in orbits are the evaporated cores of ice giants - planets that maintain their envelopes (and are therefore not Earth-like) if they form in the habitable zone. In another analysis, Neil and Rogers (2020) find evidence for two distinct populations of rocky planets, and as a result fewer Earth-sized planets in the habitable zone, for which they suggest a similar explanation. The chosen value of 7.5% is in the mid-range of values estimated by Pascucci et al. (2019) when they exclude the planets most affected by photoevaporation.

Second, we modulate the occurrence rates as a function of spectral type following Mulders et al. (2015a), who find that rocky planets are more common around lower-mass stars and tend to occupy shorter orbits. Specifically, we gradually increase the number of planets for stars less massive than the Sun and decrease their semi-major axes by interpolating between the scaling factors provided by Mulders et al. (2015a)[3] (normalized to 1 for the typical *Kepler* host star). Later, Mulders et al. (2015b) found evidence that the number of rocky planets around the typical *Kepler* M dwarf (M0 – M5) was $\sim 3.5\times$ as high as for G dwarfs, so we further increase the number of planets around M dwarfs to reflect this result. Finally, since *Kepler* was not sensitive to late M dwarfs, we assume the number of planets per star to plateau for these stars (which we believe to be a conservative extrapolation given the general trend). We note that more recent studies of M dwarf planet occurrence rates reaffirm the finding that lower-mass stars have more Earth-sized planets, including estimates from *Kepler* data (e.g., Hardegree-Ullman et al., 2019; Hsu et al., 2020) and radial velocity detections (Tuomi et al., 2019).

The net impact of these two decisions on the number of Earth-sized planets per star, as a function of stellar mass, is shown in Figure 6.2. Our estimate of $\eta_\oplus$ may seem pessimistic when compared to higher values used in predicting the detection yield of mission concepts (e.g., Gaudi et al., 2020; The LUVOIR Team, 2019; Meixner et al., 2019; Apai et al., 2019), but we view it to be a realistic estimate based on the most recent studies available. Our estimate is lower than those of Bryson et al. (2021), who avoid bias due to photoevaporation by excluding planets at high insolation. However, their resulting sample size is limited, and thus their confidence intervals are broad; indeed, our value of

---

[3]see Table 1 and Figure 4 therein

$\eta_\oplus = 7.5\%$ is within the 95% confidence interval of some of their estimates.[4]

In general, all existing estimates of $\eta_\oplus$ for G stars - including our own - are based on extrapolation and are therefore uncertain. For example, existing data cannot rule out an increase in terrestrial planet occurrence rates at orbital periods beyond ~ 100 d, which would enhance $\eta_\oplus$. To accommodate this uncertainty, we express our results in Sections 6.6 and 6.7 in terms of either $\eta_\oplus$ or the number of planets observed (which is typically linear to $\eta_\oplus$). As a result, the validity of our results is not tied to any specific value for $\eta_\oplus$.

6.3.3 Habitable zone boundaries

The circumstellar habitable zone refers to the theoretical region around a star in which a planet can sustain liquid surface water. Many formulations of the habitable zone exist, but the most commonly cited estimates are based on Kasting et al. (1993) and subsequent papers that expanded on their methodology (Kopparapu et al., 2013, 2014). In `Bioverse` we use the results of K14 to calculate the inner edge (a_{inner}, corresponding to the runaway greenhouse limit) and outer edge (a_{outer}, corresponding to the maximum greenhouse limit) of the habitable zone. To account for the dependence on planetary mass, we interpolate between the three planetary masses modeled therein.

6.3.4 Classification

Following Kopparapu et al. (2018), we classify planets as "hot", "warm", or "cold" depending on their insolation, and "rocky", "super-Earth", "sub-Neptune", "sub-Jovian", or "Jovian" depending on their size. Approximately Earth-sized planets within the habitable zone are of particular interest, as these are the most likely planets to have liquid water and habitable surface conditions. Following recent studies of detection yield estimates for direct imaging missions (Kopparapu et al., 2018; Stark et al., 2019; Gaudi et al., 2020; The LUVOIR Team, 2019), we classify as "exo-Earth candidates" (hereafter EECs) any planets with radii $0.8S^{0.25} < R < 1.4$ and orbits within the habitable zone boundaries calculated above. The lower limit on the size of EECs is the theoretical mini-

---

[4]see Table 6 therein

mum size for which a terrestrial planet can maintain an atmosphere suggested by Zahnle and Catling (2017), while the upper limit reflects the findings of several authors that planets larger than $\sim 1.4 - 1.6 R_{\oplus}$ tend to resemble mini-Neptunes in composition more than super-Earths (e.g., Weiss and Marcy, 2014; Rogers, 2015; Fulton et al., 2017).

6.3.5 Albedo and contrast ratio

Imaging measurements will be able to use a planet's brightness as a rough proxy for its size, but its brightness also depends on its geometric albedo, orbital phase, and semi-major axis. The latter two of these can feasibly be constrained by revisiting the system over several months, but it will be difficult to precisely disentangle geometric albedo and planet size. Albedo is highly sensitive to surface and atmospheric composition and will likely be highly variable for directly imaged exoplanets, so estimates of a planet's size based on brightness alone will be highly uncertain (Guimond and Cowan, 2018; Bixel and Apai, 2020a; Carrión-González et al., 2020). To properly represent this source of uncertainty, we assign geometric albedos (A_g) to each planet ranging uniformly from 10 – 70% (approximately the range of values encountered at visible wavelengths for solar system planets, e.g., Madden and Kaltenegger, 2018).

Next, we compute the planet-to-star brightness contrast ratio for each planet, modeling them as Lambertian spheres observed at quadrature phase (Traub and Oppenheimer, 2010):

$$\zeta = \frac{A_g}{\pi} \left(\frac{R_p}{a} \right)^2 \tag{6.1}$$

Note that the determination of a planet's phase from imaging data is also not trivial, requiring multiple follow-up observations to establish the orbit. Nevertheless, such observations will be a likely component of any future imaging survey in order to distinguish temperate planets from their hotter and colder peers (Gaudi et al., 2020; The LUVOIR Team, 2019).

6.3.6 Surface gravity and scale height

To translate planet radii into masses, we use the probabilistic mass-radius relationship derived by Wolfgang et al. (2016), which separates terrestrial planets and ice giants. Given each planet's mass and surface gravity, we then estimate the atmospheric scale height (h), which is important for determining the relative spectroscopic signal due to atmospheric absorption (as described in Section 6.4.3.2). We assign an atmospheric mean molecular weight μ to each planet based on its size. For "sub-Neptune" planets and larger, we assume H_2 dominated atmospheres similar to Neptune or Uranus, with $\mu = 2.5\,m_H$. For "rocky" and "super-Earth" planets, we calculate the ratio of N_2 to CO_2 based on their position relative to the habitable zone as follows. For planets within a_{inner}, we assume CO_2 dominated atmospheres similar to Venus' ($\mu = 44\,m_H$). Within the habitable zone, we adopt a positive correlation between semi-major axis and CO_2 partial pressure, which climate models predict as a result of the carbonate-silicate negative feedback mechanism (e.g., Bean et al., 2017). Specifically, we follow the correlation derived by Lehmer et al. (2020)[5], add N_2 as necessary to reach a minimum total pressure of 1 bar, and calculate the mean molecular weight between the two species ($28 < \mu < 44\,m_H$). Finally, for planets beyond a_{outer}, we assume the CO_2 to condense, leaving behind a pure N_2 atmosphere ($\mu = 28\,m_H$). We set the atmospheric temperature equal to the equilibrium temperature, assuming the Bond albedo to equal the geometric albedo. However, for EECs we assume an Earth-like atmospheric temperature due to greenhouse warming.

6.3.7 Inclination and transiting planets

Planets are assigned inclinations (i) from an isotropic distribution (i.e. a uniform distribution in $\cos(i)$ from -1 to 1). From this, and assuming circular orbits, we calculate the impact parameter on the stellar surface:

$$b = a\cos(i)/R_* \tag{6.2}$$

---

[5]We adopt the best-fit line in Figure 1 therein

For transiting planets (with $|b| < 1$) we calculate the transit depth ($\delta = (R_p/R_*)^2$) and duration:

$$T_{\text{dur}} = \frac{R_* P}{\pi a} \sqrt{1 - b^2} \tag{6.3}$$

6.3.8 Hypothetical population-level trends

The primary goal of our study is to understand which population-level trends may be detectable with a next-generation exoplanet survey. For example, could such a survey empirically determine the location of the habitable zone based on which planets have H_2O-rich atmospheres (Section 6.6), or study how oxygen evolves over time in the atmospheres of Earth-like planets (Section 6.7)?

To enable these inquiries, we apply hypothetical population-level trends to the simulated planet sample that will later be studied by simulated direct imaging and transit surveys. Specifically, we determine which planets have atmospheric water vapor based on their size and semi-major axis (following Equation 6.13), and determine which Earth-like planets have atmospheric oxygen based on their age (following Equation 6.15). A more detailed description of these assumed trends, and an assessment of their detectability by future biosignature surveys, can be found in Sections 6.6 and 6.7.

6.4 Survey simulation

The second component of `Bioverse` translates the simulated planet population from the previous section into a data set representing the result of a lengthy characterization effort with a next-generation observatory. There are a few methods by which future observatories could characterize statistically-relevant samples of habitable planets, but in `Bioverse` we focus on space-based direct imaging and transit spectroscopy. The data sets produced by these next-generation surveys will be inherently biased by the observing approach. Most notably, an imaging survey is most efficient in targeting the habitable zones of nearby FGK stars, while a transit survey is optimized for M stars. Strategic decisions also bias the data set - for example, an imaging survey must dedicate $\sim 4\times$ as much time to study a planet at 2 AU from its star versus an Earth twin, so studying planets near

the outer edge of the habitable zone will come at a steep cost.

6.4.1   Survey setup

As our template for a direct imaging survey we use LUVOIR (The LUVOIR Team, 2019, hereafter L19), a proposed NASA Flagship-class mission that would use an 8–15 meter segmented mirror and a multi-channel coronagraphic instrument to study terrestrial planets around nearby stars. While the details of the LUVOIR concept have been studied in-depth, our results are based only on its high-level characteristics - specifically, we adopt the 15-meter LUVOIR-A mirror diameter, coronagraphic inner (IWA) and outer (OWA) working angles and noise floor, and the host star catalog used to simulate its detection yield estimates (C. Stark, private correspondence). Our results should be generally applicable to any imaging mission with a similar mirror size and coronagraph.

As our template for a transit survey, we use the Nautilus Space Observatory concept (Apai et al. (2019a); Apai et al. (2019c)), which aims to study transiting exoplanets with the equivalent light-collecting area of a single 50-meter diameter telescope. To achieve this light-collecting power, Nautilus would employ an array of large telescopes with ultra-light diffractive-refractive optical elements (Milster et al., 2020) (the launch of a single, up to 8.5m diameter telescope has recently been proposed as a NASA Probe-class mission, Apai et al. (2019b)). To generate the potential list of transiting planets, we simulate systems to a distance of 150 parsecs, as our simulated surveys tend not to observe targets beyond this distance even when they are available.

Our analyses are based on a 15-meter mirror diameter imaging survey and a 50-meter diameter (equivalent area) transit survey, because among all concepts currently under consideration by the community, these are the ones purporting to offer the largest EEC sample sizes for their respective techniques. It should be noted that a 15-meter imaging survey would also be capable of characterizing nearby transiting planets as a secondary science goal, but we do not model any dual mode surveys here.

6.4.2 Which planets can be detected?

After simulating a catalog of nearby planetary systems, we discard any planets that cannot be detected by a given mission architecture. In transit mode, we exclude all non-transiting planets. In imaging mode, we exclude all planets whose maximum angular separation is less than the IWA, or whose average angular separation is greater than the OWA, or for which the planet-to-star contrast ratio (ζ) is below the instrument noise floor.

The remaining planets can, in principle, be detected by the survey, but to actually detect most of them will require preliminary observations either using the same telescope architecture or a precursor survey. A dedicated imaging mission would likely be able to detect all of the EECs that it is capable of characterizing during preliminary observations (Stark et al., 2019), but the vast majority of transiting planets within the nearest ~ 100 pc remain undiscovered. Most likely, a large-aperture spectroscopic survey of hundreds of transiting planets must be preceded by a space-based all-sky survey, similar to TESS (Ricker et al., 2015) or PLATO (Rauer et al., 2014) but with sensitivity comparable to *Kepler*. The cost and complexity of such a mission, though considerable, would likely be much less than that of a subsequent characterization effort requiring orders of magnitude greater light-collecting area.

6.4.3 Which planets can be characterized?

In-depth spectroscopic characterization is time-consuming, so the number of targets that can be characterized is a function of the total time budget allotted to the characterization effort (t_{total}). Note that t_{total} is not necessarily the same as the total survey lifetime (which might be e.g. 5–10 yr). To determine which planets can be observed within t_{total}, we first determine the amount of time required to characterize each planet, including overheads, and prioritize targets based on both their required observing time and their relative importance to the survey's goals.

6.4.3.1 Required exposure time

To determine which planets can be characterized within the time budget t_{total}, we first determine the amount of exposure time required to spectroscopically characterize a reference planet whose host star properties reflect the typical target for each survey mode. For both observing modes, the reference planet has exactly the same bulk parameters and receives the same incident flux as modern Earth. For direct imaging observations, its star is a nearby solar-type star ($T_{*,\text{ref}} = 5777$ K, $R_{*,\text{ref}} = R_\odot$, $d_{\text{ref}} = 10$ pc) while for transit observations it is a more distant early M dwarf ($T_{*,\text{ref}} = 3300$ K, $R_{*,\text{ref}} = 0.315\,R_\odot$, $d_{\text{ref}} = 50$ pc). In the examples to follow, we only consider the detection or non-detection of an absorption feature associated with a species, rather than constraints on the abundance.

We use two general circulation models (GCMs) published by Komacek and Abbot (2019) to quantify the three-dimensional atmospheric abundance profiles of our reference planets. Both models are water-covered planets around a Sun-like star (imaging mode) or early M dwarf (transit mode) with the same size, mass, and insolation as Earth and 1 bar N_2/H_2O atmospheres. These models include a treatment of ice and liquid cloud cover, which is an important factor affecting the detectability of molecular features through imaging and transit observations. Notably, because the M dwarf planet is tidally-locked, convection on its dayside is more efficient, leading to strong, high-altitude cloud cover and greater stratospheric H_2O abundance (T. Komacek, private correspondence). Finally, to enable the analysis in Section 6.7, we inject Earth's modern oxygen abundance ($pO_2 = 20.7\%$) into the model atmospheres, reducing the background N_2 pressure accordingly.

To simulate spectra for both models, we use the Planetary Spectrum Generator (hereafter PSG, Villanueva et al., 2018), which accepts three-dimensional atmospheric profiles through its GlobES module[6]. The directly imaged planet is observed at quadrature phase, while the transiting planet is observed with the night-side facing the observer. Both simulated spectra are shown in Figure 6.3 for atmospheres with and without cloud cover. Next, we use PSG to compute noise estimates for each survey architecture as a function of on-target exposure time. In imaging mode, we use the PSG template for the 15-meter

---

[6]The PSG configuration files for this study can be found in the code repository

LUVOIR-A observatory, including the projected throughput, spectral resolution, raw contrast, and detector noise for the visible and near-infrared imagers, as well as 4.5 zodis of background dust. In transit mode, we simulate observations for a 50-meter diameter aperture with 60% total throughput, ignoring detector and instrument noise. To determine whether a molecular feature can be detected, we simulate spectra with and without the target molecule and compute the detection signal-to-noise ratio (SNR) across the absorption band in a manner similar to Lustig-Yaeger et al. (2019b)[7]:

$$\text{SNR} = \sqrt{\sum_i (\Delta y_i / \sigma_{y_i})^2} \tag{6.4}$$

where Δy_i is the difference between the two spectra in each spectral bin and σ_{y_i} is the measurement uncertainty. Finally, we compute the exposure time required to achieve a SNR = 5 detection of the feature for the reference planet (t_{ref}) in each survey mode, then scale this value to determine the exposure time required for each individual planet detected by the survey.

6.4.3.2 Exposure time scaling

We define t_i as the amount of exposure time required to spectroscopically characterize a planet at wavelength λ_{eff}. If we assume that t_i depends primarily on the number of photons collected, then we can estimate it by scaling t_{ref} (as determined using PSG) as follows:

$$\frac{t_i}{t_{\text{ref}}} = f_i \left(\frac{d_i}{d_{\text{ref}}} \right)^2 \left(\frac{R_*}{R_{*,\text{ref}}} \right)^{-2} \left(\frac{B_{*,i}(\lambda_{\text{eff}}, T_{*,i})}{B_{\odot}(\lambda_{\text{eff}}, T_{*,\text{ref}})} \right)^{-1} \tag{6.5}$$

where f_i summarizes the factors affecting the signal strength unique to each observing mode. In imaging mode, the exposure time is inversely proportional to the planet-to-star contrast ratio (assuming observations at quadrature phase):

$$f_i^{\text{im}} = \left(\frac{\zeta_i}{\zeta_{\oplus}} \right)^{-1} \tag{6.6}$$

In transit mode, the transit depth signal induced by the atmosphere is (to first order) $\Delta \delta \sim (R_p/R_*)^2 (h/R_p)$ (Winn, 2010) and the required exposure time is inversely proportional to

---

[7]Equations 4–6 therein
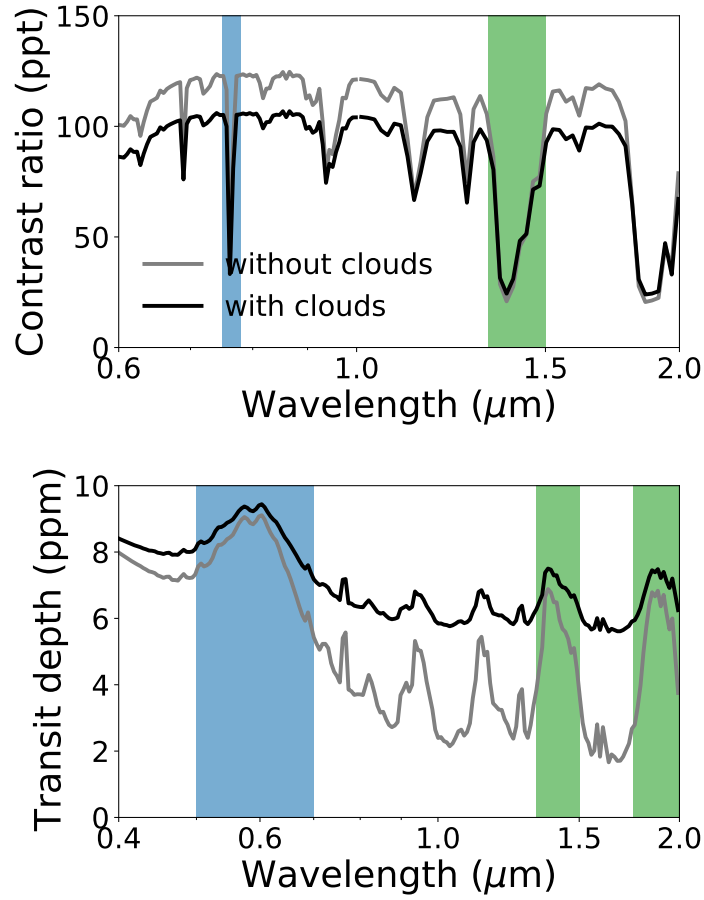
Figure 6.3: Model spectra for the reference planet in imaging (top; contrast ratio in parts-per-trillion) and transit (bottom; transit depth in parts-per-million) mode. The spectra are based on GCM models published by Komacek and Abbot (2019), who investigate ice and liquid cloud cover on planets as a function of spectral type and tidal locking. We include the effects of clouds to determine our exposure time estimates (black), while clear-sky spectra are shown for reference (gray). Targeted absorption bands include H_2O (green) and O_2 or O_3 (blue).

its square:

$$f_i^{\mathrm{tr}} = \left(\frac{h_i}{h_\oplus}\right)^{-2} \left(\frac{R_{p,i}}{R_\oplus}\right)^{-2} \left(\frac{R_{*,i}}{R_{*,\mathrm{ref}}}\right)^4 \tag{6.7}$$

We round up t_i to the next integer multiple of the planet's transit duration, because a transit survey would likely observe complete transits to measure the baseline. Planets are considered to be invalid targets if the total number of required transit observations is greater than either the number of available transits within 10 years or 10^3.

These scaling relations are meant to capture the main factors affecting the relative exposure time required for each target so as to provide an approximate mapping between the total amount of time dedicated to a survey and the number and distribution of targets it can observe. Ultimately, the primary metric affecting a survey's statistical power is usually the number of EECs characterized, and we translate t_{total} into the number of characterized EECs so the reader can interpret our results as a function of sample size.

6.4.3.3 Overheads

In imaging mode, following L19 we increment each planet's required exposure time by 2 hr to account for slew overheads and overheads associated with wavefront control. These overheads end up being relatively insignificant except for the closest targets. In transit mode, we assume 0.5 hr of slew overheads per observation, plus a total overhead equal to the transit duration for baseline observations before and after each transit event.

6.4.3.4 Target prioritization

Given a limited time budget, it seems reasonable to prioritize observations of planets in order of increasing t_i so as to maximize the number of planets observed. However, prioritizing targets strictly by t_i will lead to a biased sample, especially in the case of transit surveys that are strongly biased towards the detection of close-in planets. To counter-act these biases, we assign a weight w_i to each planet, and calculate its priority as follows:

$$p_i = w_i/t_i \tag{6.8}$$

The specific choice of w_i depends on the hypothesis being tested and is discussed in Sections 6.6 and 6.7. To create the final simulated data set, we observe targets in order of decreasing p_i until some pre-determined time limit t_{total} is reached.

6.4.4 Comparison between survey modes

In the following sections, we use `Bioverse` to evaluate the statistical potential of direct imaging and transit spectroscopy surveys, but we avoid direct comparisons of their results for the following reasons. First, the technical requirements for and limitations of a direct imaging biosignature survey have been more thoroughly explored due to investments in the LUVOIR and HabEx mission concepts. As a result, our results for the transit survey are likely more optimistic. Second, we do not wish to imply that a survey's statistical power is the only or most important dimension for comparison, as each architecture enables unique capabilities that the other does not.

For the topics discussed here, the primary difference between the two surveys is the number of EECs each can characterize. For the 15-meter imaging survey, this number is 15–20, and is volume- rather than time-limited. This estimate is consistent with that of L19 when adjusted for our updated value of $\eta_\oplus$ ($\approx 7.5\%$ for G stars). For the 50-meter (equivalent area) transit survey, this number grows with time, with e.g. 60–70 EECs being surveyed for H_2O absorption or ~ 200 for O_3 absorption given $t_{total} = 2$ yr.

6.5 Hypothesis testing

The third component of `Bioverse` assesses the information content contained within the simulated data sets from the previous section. This assessment focuses on two primary questions: first, how likely is it that the survey would be able to detect the effects of a statistical trend injected into the simulated planet population (Section 6.3.8)? Second, how precisely could the survey constrain the parameters of that trend? To answer these questions, we rely on a standard Bayesian hypothesis testing approach.[8]

---

[8]For a review of Bayesian parameter estimation and model selection in astronomy, we refer the reader to Trotta (2008).

6.5.1 Null and alternative hypotheses

Each simulated data set can be thought of as a set of independent variables x and dependent variables y. For this section (and the examples to follow), we consider x and y to each represent measurements of a single variable, but this hypothesis testing framework can extend to multivariate measurements as well. The hypothesis $h(\vec{\theta}, x)$ describes the relationship between the x and y in terms of a set of parameters $\vec{\theta}$. The simplest hypothesis is the null hypothesis, in which there is no relationship:

$$h_{\mathrm{null}}(\theta, x) = \theta$$

The null hypothesis is compared to an alternative hypothesis, which proposes a relationship between x and y, using a Bayesian parameter estimation and hypothesis testing approach.

6.5.2 Likelihood function and prior distribution

Given a hypothesis h, the likelihood function takes on one of two forms. In the case where y is binary (e.g., the detection or non-detection of an atmospheric species), then h is the probability that $y = 1$, and the likelihood function is:

$$\mathcal{L}(y|\vec{\theta}) = \prod_{i}^{N} \left[y_i h(\vec{\theta}, x_i) + (1 - y_i)(1 - h(\vec{\theta}, x_i)) \right] \tag{6.9}$$

Alternatively, if y is a continuous variable measured with normal uncertainty σ_y, then h predicts the expectation value of y, and the likelihood is described by the normal distribution:

$$\mathcal{L}(y|\vec{\theta}) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi\sigma_{y,i}^2}} \exp\left(-\frac{(y_i - h(\vec{\theta}, x_i))^2}{2\sigma_{y,i}^2} \right) \tag{6.10}$$

Note that in both example applications of `Bioverse` to follow, we consider a detection or non-detection as our dependent variable and use the likelihood function defined by Equation 6.9.

The parameter prior distribution is denoted by $\Pi(\vec{\theta})$. Given limited prior information about the true values of parameters $\vec{\theta}$, we generally assume uniform or log-uniform dis-

tributions spanning the range of plausible values. Further justification for our choice of prior distributions can be found in the examples to follow.

6.5.3 Parameter estimation and Bayesian evidence

For each simulated data set, we sample the posterior distribution of the hypothesis parameters $\vec{\theta}$ using a Markov Chaint Monte Carlo (MCMC) algorithm, implemented by `emcee` (Foreman-Mackey et al., 2013). This sampling yields measurement constraints of the parameters $\vec{\theta}$. We also use a nested sampling algorithm (Skilling, 2006), implemented by `dynesty` (Speagle, 2020), to estimate the Bayesian evidence for the alternative hypothesis:

$$\mathcal{Z} = P(y|h) = \int \mathcal{L}(y|\vec{\theta})\Pi(\vec{\theta})d\theta \tag{6.11}$$

To test a hypothesis, we can compare its evidence to that of the null hypothesis, finding evidence to reject the null hypothesis when:

$$\Delta\ln(\mathcal{Z}) = \ln(\mathcal{Z}) - \ln(\mathcal{Z}_{\text{null}}) > 3 \tag{6.12}$$

We choose $\Delta\ln(\mathcal{Z}) > 3$ as our threshold because it corresponds to the common $p < 0.05$ threshold for hypothesis testing with other frequentist tests (e.g., Student's t-test).

It should be noted that `dynesty` also samples the parameter posterior distributions - so why use `emcee` to do this separately? In short, nested sampling is optimized to measure $\mathcal{Z}$, while MCMC is optimized to determine the posterior distribution. While `dynesty` can quickly compute the Bayesian evidence with sufficient accuracy ($\sigma_{\ln(\mathcal{Z})} \lesssim$ 0.5), we find it takes significantly longer to converge to the same parameter posterior distributions as `emcee`. Since we repeat each simulated survey $> 100,000$ times, we find this mixed approach to be necessary to achieve both accurate evidence and parameter estimations on a reasonable timescale.

6.5.4 Statistical power

Whether or not an individual simulated survey is able to reject the null hypothesis can often depend on stochastic error; one simulated survey may be able to reject the null

hypothesis where another cannot. To summarize our results, we re-run each simulated survey several times under the same set of assumptions and calculate the fraction of survey realizations that achieve a positive result. This metric is also known as the statistical power, and it allows us to assess a survey's statistical potential as a function of both survey parameters (such as total survey duration) and as-yet unknown astrophysical parameters (such as the frequency of habitable planets).

This concludes the description of the three primary components of `Bioverse`. In the following two sections, we will demonstrate applications of `Bioverse` to its stated goal of assessing the statistical power of next-generation biosignature surveys.

6.6   Example 1: Empirical determination of the habitable zone boundaries

Models of the habitable zone predict that planets with oceans can only exist within a finite - and perhaps very narrow - range of insolations. An associated prediction is that terrestrial planets in the habitable zone with water-rich atmospheres are the most likely candidates for ocean-bearing worlds. These models will play an important role in the design and target prioritization of next-generation observations; for example, preliminary search strategies for future biosignature surveys often dedicate intensive follow-up to water-bearing habitable zone planets (The LUVOIR Team, 2019), while delegating non-habitable zone planets to a lower priority. However, models for the habitable zone have not been tested outside of the solar system, and estimates of its location and width have varied by factors of several over the past few decades.

Could future observatories use data acquired from preliminary observations to test the "habitable zone hypothesis" i.e., the hypothesis that planets with water vapor should be more abundant within a narrow and finite range of orbital separations? Further, could these data be used to empirically determine the location and width of the habitable zone? The practical benefit of testing the habitable zone hypothesis would be to make the survey's target prioritization strategy more efficient and to better determine which of its

targeted planets are most likely to be habitable. By measuring its boundaries, observers could test the predictions of various habitable zone models, and therefore the physical mechanisms on which they rely. Finally, empirical constraints on the width of the habitable zone will be important for determining the occurrence rate of habitable worlds. Here, we use `Bioverse` to explore how a survey of atmospheric water vapor could be used to test the habitable zone hypothesis.

6.6.1   Model predictions

Climate models predict a steep decline in water vapor abundance of terrestrial planets outside of the habitable zone. Within the inner edge, an Earth-like planet may undergo a runaway greenhouse as on Venus, leaving behind only a tenuous amount of atmospheric water vapor. Beyond the outer edge, the oceans may freeze, and water vapor would not accumulate except in very low pressure atmospheres that permit its sublimation.

In `Bioverse` we implement these predictions as follows. We assume that a fraction $f_{\text{EEC}}^{\text{H}_2\text{O}}$ of EECs are in fact habitable, meaning they bear surface water and atmospheric water vapor. We also allow a fraction $f_{\text{non-EEC}}^{\text{H}_2\text{O}}$ of non-EECs to have atmospheric water vapor, serving as a source of noise and "false positives" for habitable planets. Then the fraction of planets with atmospheric water vapor can be described as:

$$
f^{\text{H}_2\text{O}} = \begin{cases} f_{\text{EEC}}^{\text{H}_2\text{O}} & \text{if } a_{\text{inner}} < a < a_{\text{outer}} \\[4pt] & \text{and } 0.8S^{0.25} < R < 1.4R_{\oplus} \\[4pt] f_{\text{non-EEC}}^{\text{H}_2\text{O}} & \text{if } a < a_{\text{inner}} \text{ or } a > a_{\text{outer}} \\[4pt] & \text{and } R > 0.8S^{0.25} \\[4pt] 0 & \text{if } R < 0.8S^{0.25} \end{cases}
\tag{6.13}
$$

where the habitable zone boundaries and planet size limits are those discussed in Sections 6.3.3 and 6.3.4.
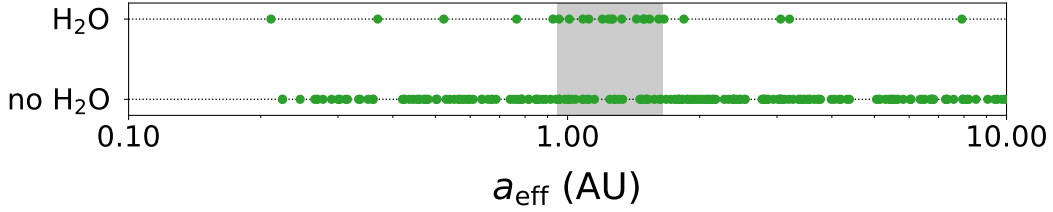
Figure 6.4: An example of a simulated direct imaging data set for Section 6.6. Planets are probed for the presence of atmospheric water vapor across a broad range of orbital separations. We assume the habitable zone (gray) to be marked by an abundance of water-rich atmospheres. The separation $a_{\text{eff}} = a(L_*/L_\odot)^{-1/2}$ is the solar-equivalent semi-major axis.

6.6.2 Simulated survey

6.6.2.1 Measurements

The imaging and transit surveys perform a set of measurements outlined in Table 6.2 to determine the size and orbital separation of each potential target. In imaging mode, the planet's size is not determinable without prior knowledge of the geometric albedo, so an estimated size (R_{est}) that assumes Earth-like reflectivity is used as a proxy. In both modes, the orbital separation is converted to the "effective" semi-major axis (a_{eff}) for which the planet would receive the same insolation around a Sun-like star.

These preliminary measurements are used to prioritize targets as discussed in the following section. Those targets of high enough priority are spectroscopically characterized to determine whether their atmospheres contain H_2O. The final output of each simulated survey as a data set consisting of (a_{eff}, H_2O), where $H_2O = \{0, 1\}$ reflects the absence or presence of water absorption features in the planet's spectrum. One example of a simulated data set is shown in Figure 6.4.

Table 6.2: Measurements made by the simulated surveys in Examples 1 and 2. Parameters marked by † are calculated from other measured values.

Parameter	Measurement uncertainty	Description / notes
Example 1		
Imaging survey		
L_*	negligible	Host star luminosity
ζ	15%	Planet-to-star contrast
a	10%	Semi-major axis
a_{eff}†	10%	Solar-equivalent semi-major axis
R_{est}†	10%	Estimated radius assuming Earth-like reflectivity
H_2O	Detected / not detected	Presence of 1.4 μm H_2O absorption
Transit survey		
M_*, R_*	5%	Host star mass and radius
P	negligible	Orbital period
δ	negligible	Baseline transit depth
a_{eff}†	1.7%	Solar-equivalent semi-major axis
R†	5%	Planet radius
H_2O	Detected / not detected	Presence of 1.4 μm and 1.9 μm H_2O absorption

Table 6.2: (continued)

Parameter	Measurement uncertainty	Description / notes
Example 2		
Imaging survey		
t_*	10%	Age (as measured through astero-seismology)
O_2	Detected / not detected	Presence of 0.7 μm O_2 absorption
Transit survey		
t_*	30%	Age (model-based estimate)
O_3	Detected / not detected	Presence of 0.6 μm O_3 absorption

6.6.2.2 Target prioritization

To test the habitable zone hypothesis we must observe planets spanning a broad range of semi-major axes, but prioritizing targets solely based on required exposure time will bias observations towards close-in planets. Furthermore, planets much smaller or larger than Earth are not likely to be habitable regardless of insolation, and therefore serve as a source of noise. The counter these effects, we weight each target according to its size and orbital separation following Figure 6.5. We tuned this prioritization based on trial and error to achieve the following goals:

1. Prioritize observations of more probable Earth analogs (planets receiving 50–150% of Earth's incident flux).

2. Balance observations of widely-separated planets versus close-in planets.

3. Minimize observations of non-Earth sized planets.

In transit mode, we additionally weight each target by (a/R_*) to negate the bias due to close-in planets being more likely to transit. The resulting distribution of observed planets is also shown in Figure 6.5.
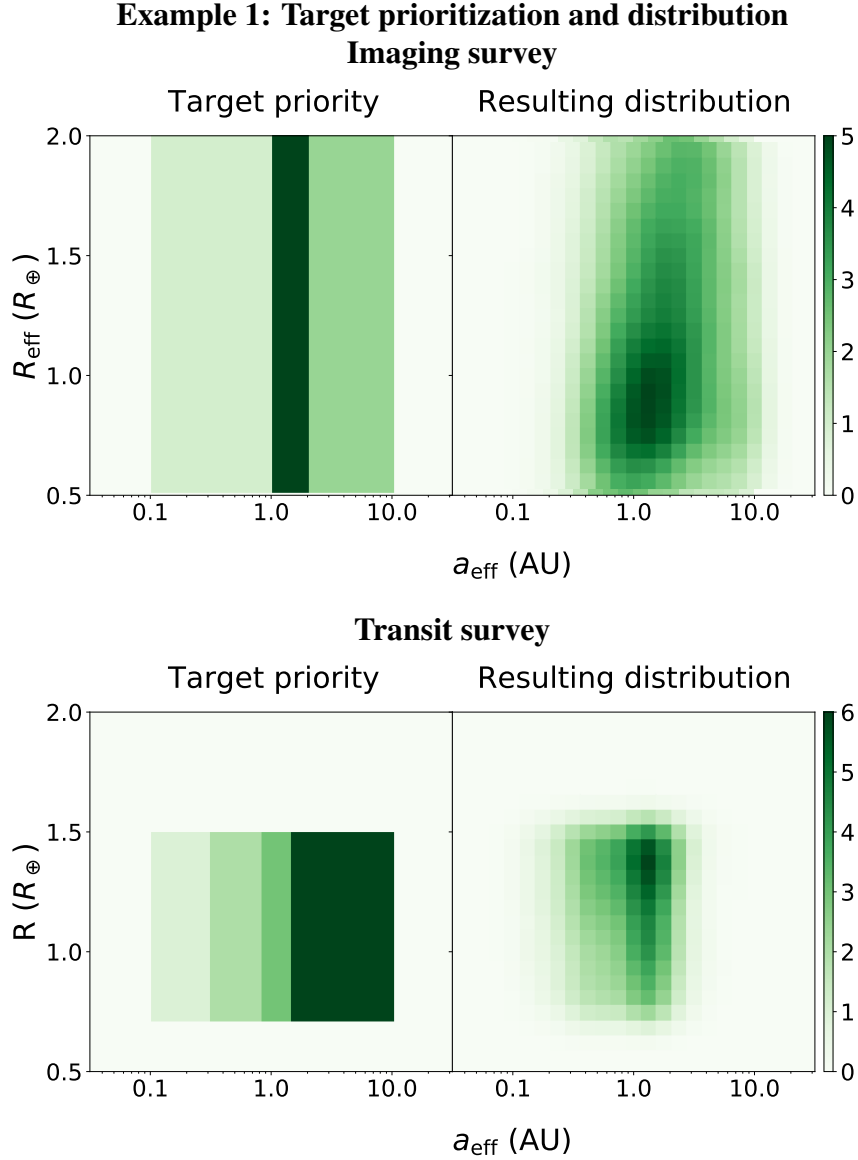
Figure 6.5: Summary of target prioritization for the simulated imaging (top) and transit (bottom) surveys in Section 6.6. The left panel shows the relative weight assigned to each target as a function of size and orbital separation (w_i in Equation 6.8). The right panel shows the resulting relative distribution of targets that can be probed for the presence of water vapor within the survey duration. In the case of the imaging survey, the planet size cannot be directly measured, so the "estimated" radius (assuming Earth-like reflectivity) is used as a proxy. In the case of the transit survey, an additional weight is applied to counteract the R_*/a transit probability (not shown above).

Table 6.3.   Parameter Prior Distributions for Equations 6.14 and 6.16.

Parameter	Description	Prior limits (log-uniform distribution)
Example 1		
a_{inner}	Inner edge of the habitable zone	0.1 – 2.0 AU
Δa	Width of the habitable zone	0.01 – 10 AU
f_{HZ}	Fraction of habitable zone planets with H_2O	0.001 – 1
$(f_{\mathrm{non\text{-}HZ}}/f_{\mathrm{HZ}})$	Fraction of non-habitable zone planets with H_2O (relative to f_{HZ})	0.001 – 1
Example 2		
f_{life}	Fraction of EECs with life	0.001 – 1
$t_{1/2}$	Oxygenation timescale of inhabited planets	0.1 to 100 Gyr

Table 6.4.   Predicted Signal Strengths of H_2O (Example 1) and O_2 or O_3 (Example 2) Absorption

Survey mode	Feature	Wavelength	Signal strength (without clouds)	Signal strength (with clouds)	Time required (hr) (with clouds)
Imaging	H_2O	1.4 μm	90 ppt	55 ppt	0.9
	O_2	0.76 μm	90 ppt	70 ppt	2.6
Transit	H_2O	1.4 μm	3.5 ppm	0.5 ppm	181
		1.9 μm	5 ppm	0.7 ppm	
	O_3	0.6 μm	4 ppm	2 ppm	74

Note. — Values are calculated for the representative target of each survey mode. Signal strengths are expressed as the peak amplitude of the change in planet-to-star contrast ratio (in parts-per-trillion) or transit depth (in parts-per-million) within the absorption band. The exposure time required for a 5σ detection is determined using PSG, and scaled for each individual target according to Equation 6.5. In transit mode, the signals from two bands are combined to achieve the detection of H_2O. In imaging mode, we select the feature that requires the least exposure time to detect.

6.6.2.3 Time budget

Following the procedure in Section 6.4.3.1, we use PSG to determine the exposure time required for a 5σ detection of water vapor absorption through its near-infrared absorption bands. In transit mode, we combine the SNR from the 1.4 and 1.9 μm features. In imaging mode, we only target the 1.4 μm band, as LUVOIR will be unable to observe the full near-infrared spectrum simultaneously, and the 1.9 μm band is harder to observe due primarily to lower stellar flux.

In imaging mode, we find $t_{\mathrm{ref}} = 0.9$ hr are necessary to probe the reference planet for water vapor, while in transit mode we require $t_{\mathrm{ref}} = 181$ hr of in-transit exposure time; the details of these calculations are shown in Table 6.4. Next, we scale t_{ref} according to Equations 6.5 through 6.7 to determine the amount of exposure time required to characterize each planet, then add observing overheads. We weight the targets according to Figure 6.5, calculate each target's priority following Equation 6.8, then finally observe by order of decreasing probability until the total time budget t_{total} is reached.

The average number of EECs observed by each survey is displayed in Figures 6.6a and 6.7a as a function of either $\eta_{\oplus}$ or t_{total}. While we use the number of EECs observed as our primary metric of sample size, note that most observed targets are non-habitable. Since the imaging survey yield quickly becomes volume-limited, we investigate the impact of varying $\eta_{\oplus}$ for a fixed time budget $t_{\mathrm{total}} = 120$ d (which is sufficient to characterize $> 90\%$ of detectable EECs). For the transit survey, we fix $\eta_{\oplus} = 7.5\%$ and investigate the impact of varying t_{total}.

6.6.3 Habitable zone hypothesis

Now, let us approach the simulated data from the view of an observer who no prior knowledge of Equation 6.13 using the Bayesian hypothesis testing framework outlined in Section 6.5. The habitable zone hypothesis states that planets within the habitable zone are more likely to have water vapor than those outside of it:

$$h^{\mathrm{HZ}}(a_{\mathrm{eff}}) = \begin{cases} f_{\mathrm{HZ}} & \text{if } a_{\mathrm{inner}} < a_{\mathrm{eff}} < a_{\mathrm{inner}} + \Delta a \\ f_{\mathrm{HZ}}(f_{\mathrm{non\text{-}HZ}}/f_{\mathrm{HZ}}) & \text{otherwise} \end{cases} \tag{6.14}$$

This is a four parameter model with $\vec{\theta} = [a_{\mathrm{inner}}, \Delta a, f_{\mathrm{HZ}}, (f_{\mathrm{non\text{-}HZ}}/f_{\mathrm{HZ}})]$. The choice of parameters was driven by two factors: first, the width of the habitable zone (Δa) is relevant for testing "rare Earth" models in which the habitable zone is very narrow. Second, we can use simple log-uniform prior distributions for these parameters without having to filter out parameter combinations that violate the assumptions of the habitable zone hypothesis (e.g., $f_{\mathrm{non\text{-}HZ}} > f_{\mathrm{HZ}}$).

6.6.4 Prior assumptions

K14's model for the habitable zone, which we implement in the simulated planet population, assumes a carbon-silicate feedback cycle that enhances CO_2 concentrations for planets further from their host stars, and spans 0.95–1.67 AU for the Sun. While this estimate has strong heritage (Kasting et al., 1993; Kopparapu et al., 2013), it has also been preceded and succeeded by more conservative or generous estimates, which we use to set the prior distribution of values considered for a_{inner} and Δa.

Estimates of the inner edge range as far inward as 0.38 AU (for highly-reflective desert worlds with a minimal greenhouse effect, Zsom et al., 2013) and we allow that the inner edge could be as far out as 2 AU, in which case Earth would be an unusually cool outlier. Estimates of the habitable zone width have varied as well; a classic estimate by Hart (1979) suggests a very narrow habitable zone ($\Delta a < 0.1$ AU), which would imply that Earth-like planets are especially rare. More recent estimates have proposed mechanisms by which the habitable zone could extend as far as 2.4 AU (Ramirez and Kaltenegger, 2017) or even 10 AU (Pierrehumbert and Gaidos, 2011) from the Sun. Given this wide range of estimates for a_{inner} and Δa, we assume broad prior distributions for both, shown in Table 6.3.

6.6.5 Results

We repeat the simulated survey and Bayesian analysis $> 10{,}000$ times over a grid of values for the astrophysical parameters in Equation 6.13 for both survey architectures. With each simulated survey, we use `dynesty` to calculate the Bayesian evidence in favor

Example 1: Results for 15-meter imaging survey
How many exo-Earth candidates are probed for H$_2$O?

Can this survey test the habitable zone hypothesis?

Statistical power   Required EEC yield

Can this survey determine the location of the habitable zone?
(results for six random realizations of the survey)
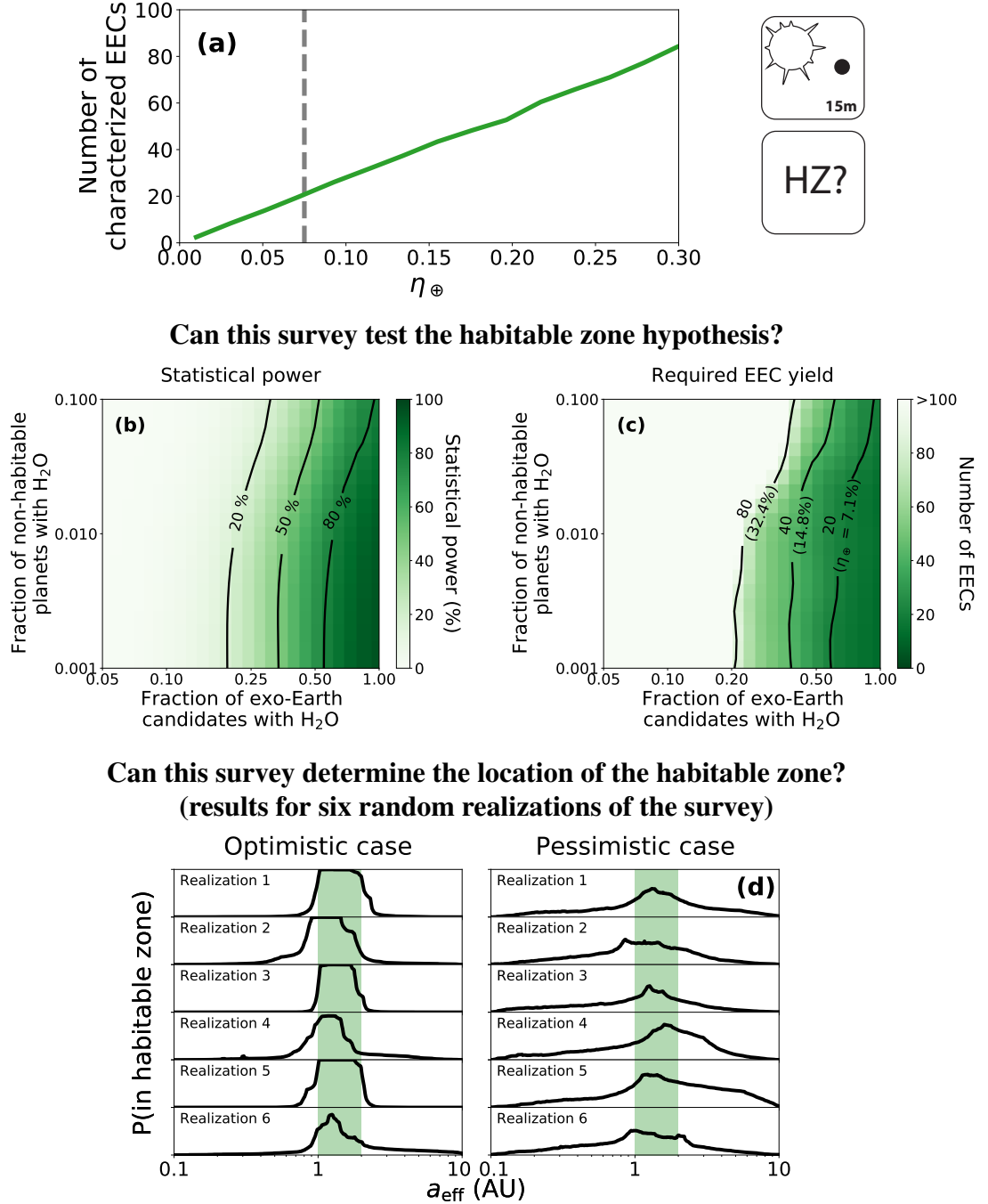
Optimistic case   Pessimistic case

Figure 6.6: Results for the imaging survey in Section 6.6. (a) The number of EECs observed versus $\eta_\oplus$ (for G stars), assuming $t_{\text{total}} = 120$ d. As our baseline case, we set $\eta_\oplus = 7.5\%$. (b) The statistical power to test the habitable zone hypothesis as a function of the astrophysical parameters in Equation 6.13.

Figure 6.6: (cont.) (c) The minimum number of EECs that must be characterized to achieve 80% statistical power, with the corresponding values of $\eta_\oplus$. (d) The posterior probability that a planet with effective separation a_eff is in the habitable zone, as estimated by six random realizations of the survey under an optimistic case (80% of EECs are habitable, left) and pessimistic case (20% of EECs are habitable, right). The true habitable zone is highlighted in green, and in both cases 1% of non-habitable planets have H_2O.

of the habitable zone hypothesis, and `emcee` to sample the posterior distributions of a_inner and Δa. The results are summarized by Figure 6.6 and 6.7 for the simulated imaging and transit surveys, respectively.

6.6.5.1 Imaging survey

An ambitious direct imaging survey with a 15-meter telescope could confidently detect the habitable zone with a 3-month long observing campaign provided most EECs are habitable. If habitable planets are less common, however, then more EECs must be observed. The EEC yield of an imaging survey is typically volume-limited, so higher values of $\eta_\oplus$ would be required to test this hypothesis for more pessimistic astrophysical parameters. In the best case scenario ($\eta_\oplus \approx 40\%$), a 15-meter imaging mission could perform the test if 20% of EECs are habitable, but this value for $\eta_\oplus$ is likely too optimistic.

If $\sim 80\%$ of EECs are habitable, the imaging survey would be able to measure the location of the habitable zone with sufficient accuracy to exclude some more extreme estimates of its boundaries with reasonable confidence. In particular, it would be able to place a confident lower bound on Δa, rejecting some "rare Earth" models, which predict a very narrow habitable zone (e.g., Hart, 1979).

Finally, it should be noted that imaging surveys will have access to planet brightness and color information that could be incorporated into this analysis; for example, albedo and photometric color may vary predictably across the habitable zone (Checlair et al., 2019). Hypotheses that include this information could be tested with better statistical power and parameter constraints than the one examined here.

**Example 1: Results for 50-meter (equivalent area) transit survey
How many exo-Earth candidates are probed for H$_2$O?**
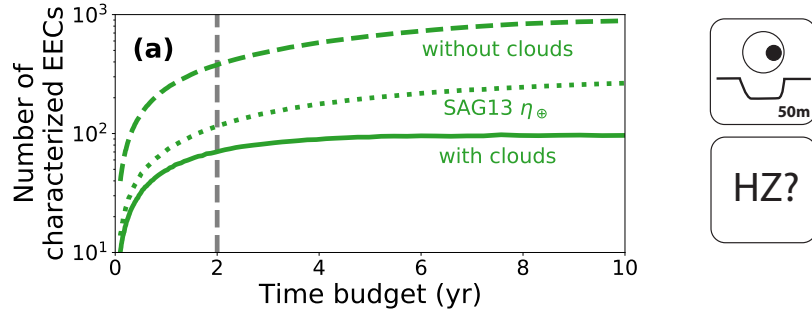
Can this survey test the habitable zone hypothesis?
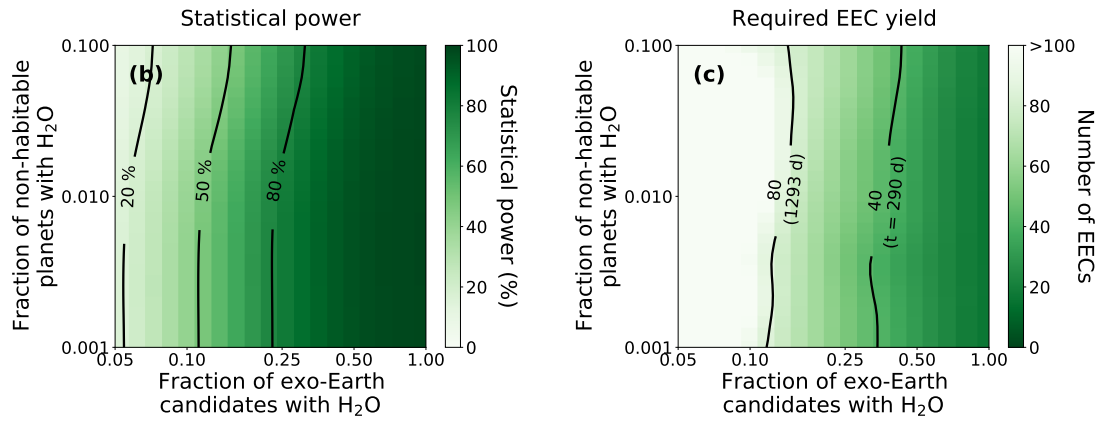
**Can this survey determine the location of the habitable zone?
(results for six random realizations of the survey)**
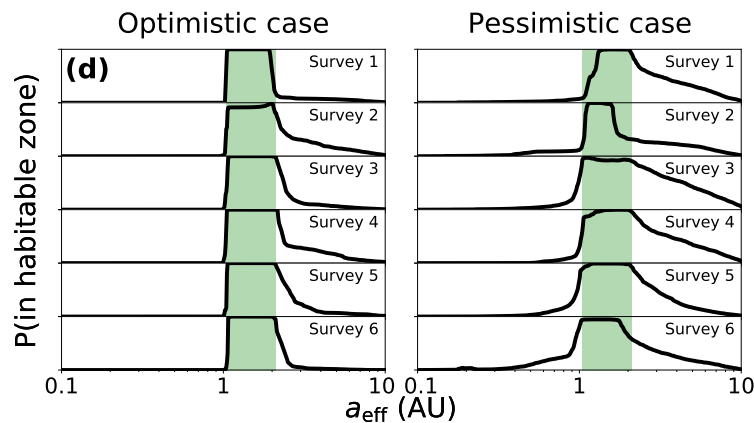
Figure 6.7: Results for the transit survey in Section 6.6. (a) The number of EECs observed versus the observing time budget, assuming $\eta_\oplus = 7.5\%$ for G stars and cloudy atmospheres (solid). 4–10× as many planets could be observed if clouds were neglected (dashed), or 1.5–3× as many with clouds if assuming the higher SAG13 estimate of $\eta_\oplus = 24\%$ (dotted). As our baseline case, we set $t_{\text{total}} = 2$ yr.

Figure 6.7: (cont.) (b) The statistical power to test the habitable zone hypothesis as a function of the astrophysical parameters in Equation 6.13. (c) The minimum number of EECs that must be characterized to achieve 80% statistical power, with the necessary observing time budget t_{total}. (d) The posterior probability that a planet with effective separation a_{eff} is in the habitable zone, as estimated by six random realizations of the survey under an optimistic case (80% of EECs are habitable, left) and pessimistic case (20% of EECs are habitable, right). The true habitable zone is highlighted in green, and in both cases 1% of non-habitable planets have H_2O.

6.6.5.2 Transit survey

The transit survey can confidently detect the habitable zone even in the case where most EECs are not habitable ($f_{\text{EEC}}^{H_2O} \approx 25\%$), provided 60–70 EECs can be probed for atmospheric water vapor during a 2-year characterization effort. Furthermore, the full survey duration may not be necessary if most EECs end up to be habitable (in which case a shorter 3–6 month survey would suffice).

The transit survey can precisely measure the inner edge of the habitable zone to within ± 0.1 AU of its true location in most simulated surveys if most EECs are habitable, and can sometimes accomplish this even if most EECs are not habitable. The width (or outer edge) is more difficult to constrain as the transit survey only observes a handful of planets beyond the outer edge (~ 10). This bias has two causes: first, planets outside of the habitable zone are less likely to transit, so they are typically found around more distant stars. Second, colder planets have smaller atmospheric scale heights, and therefore weaker H_2O absorption features. Both of these effects increase the time required to characterize cold planets, making them low priority targets.

6.6.6 Discussion

6.6.6.1 Impact of clouds

Clouds will have a major impact on the transit survey's ability to test the habitable zone hypothesis, as they dampen the absorption signal due to tropospheric water vapor and therefore increase the number of transit observations required to detect it. As shown in Figure 6.7a, this means that a much smaller number of targets can be observed within

a fixed time budget, and many of the most distant targets become infeasible to characterize as it would require the combination of decades' worth of transit observations. A possible mitigating strategy would be to expand the observatory's light-collecting area. The Nautilus Space Observatory, on which we base our transit survey results (Apai et al., 2019), would consist of 35 identically-manufactured unit telescopes. As such, the cost would scale linearly with light-collecting area, and doubling the number of telescopes would reduce by nearly half the number of transit observations required to characterize each planet.

Our cloud assumptions are based on the GCM models of Komacek and Abbot (2019), who show that tidally locked planets around M dwarfs have much higher dayside cloud covering fractions than Earth-like planets. If this bears true, it will likely prevent the characterization of such planets through transit spectroscopy by JWST (Fauchez et al., 2019; Komacek et al., 2020; Suissa et al., 2020; Pidhorodetska et al., 2020) and possibly even larger observatories. In the pessimistic case, even a 50-meter equivalent area transit survey may be unable to detect atmospheric water vapor for all but a handful of nearby exo-Earths orbiting M dwarfs, so the survey must target more distant K and G dwarfs instead. This will come at the cost of sample size, as we estimate that the increased average distance, less frequent transits, and lower transit depths for habitable zone planets around these stars will outweigh their higher stellar luminosity in terms of observing time cost.

Clouds impact imaging observations as well, although they have little effect on the results presented here. We expect cloud cover to be less prevalent for non-tidally locked planets orbiting Sun-like stars, and highly-reflective clouds at low enough altitudes can have a beneficial effect on imaging observations as they amplify the absorption due to molecules in higher layers. More importantly, the imaging survey modeled here is volume- rather than time-limited, so with or without clouds we find that the survey can probe its entire EEC sample for water absorption within less than three months.

Our exposure time and sample size estimates for the transit survey are based on one possible realization of cloud conditions, but cloud cover may vary greatly across targets and observation epochs. Indeed, in the GCM models we employ, the effect of clouds

on transmission spectra is sensitive to orbital period, spectral type, cloud particle size, and many other parameters (Komacek et al., 2020), suggesting that the actual distribution of cloud properties in terrestrial exoplanet atmospheres may be fairly broad. An efficient transit survey could seek to identify planets with clearer atmospheres (e.g. through scattering features in visible light) and prioritize these over cloudier targets, thereby increasing the sample size.

6.6.6.2 Effect of non-habitable H_2O-rich atmospheres

Naturally, the habitable zone hypothesis is easier to test if more habitable planets are observed, and the number of EEC characterizations required to test it is approximately proportional to the fraction of EECs that are habitable ($f_{\mathrm{EEC}}^{\mathrm{H_2O}}$). However, non-habitable planets are far more common than habitable planets, so if even a small fraction ($f_{\mathrm{non\text{-}EEC}}^{\mathrm{H_2O}}$) of these have H_2O, the statistical excess of H_2O in the habitable zone will be muted. In general, we find the statistical power to be unaffected provided that $f_{\mathrm{non\text{-}EEC}}^{\mathrm{H_2O}} \lesssim 1\%$, but the impact can be considerable if $f_{\mathrm{non\text{-}EEC}}^{\mathrm{H_2O}} \gtrsim 10\%$. This result seems sensible, as approximately 10% of the total sample are EECs, so $f_{\mathrm{non\text{-}EEC}}^{\mathrm{H_2O}} > 10\%$ would imply that H_2O-rich non-habitable planets are more common than habitable planets.

Our assumption in Equation 6.13 is that all EECs with water vapor are habitable, and the fraction of non-EECs with water vapor is mostly independent of insolation. However, if such "false positives" exist, their abundance is likely a function of insolation. For example, consider a population of non-habitable planets whose surfaces have been desiccated by a runaway greenhouse effect but that still maintain thick, H_2O-rich atmospheres. Such planets should be clustered near the inner edge of the habitable zone (e.g., Turbet et al., 2019), appearing as an extension of the habitable planet population to high insolations rather than as a distinct planet population. Even planets defined as EECs may actually be non-habitable (due to differences in initial volatile content, plate tectonics, outgassing rates, etc.) yet still possess water vapor, making them statistically indistinguishable from habitable EECs. Again, the effect of these false positives will likely be negligible provided they are much less common than habitable planets, but indicators of planetary (non-)habitability other than H_2O may be necessary to filter them out.
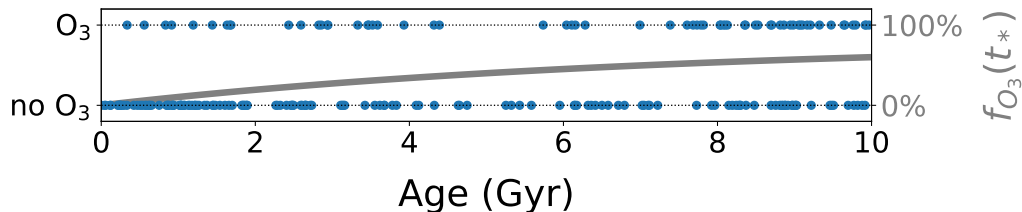
Figure 6.8: An example of a simulated transit spectroscopy data set for Section 6.7. Earth-sized planets in the habitable zone are probed for the presence of O_3 (a tracer of O_2), which we assume becomes more common with age as more planets undergo global oxidation events. This "age-oxygen correlation" (Equation 6.16) is represented by the grey line, in this case where $f_{\text{life}} = 80\%$ of observed planets are inhabited and the oxygenation timescale is 5 Gyr. Age estimates are uncertain to $\pm 30\%$.

6.7 Example 2: Evolution of Earth-like Planets

By characterizing a sufficiently-large sample of terrestrial worlds, a next-generation observatory could test hypotheses for how they evolve over time. One such hypothesis is that inhabited planets with oxygen-producing life, like Earth, evolve towards greater oxygen content over Gyr timescales due to long-term changes in global redox balance. As we propose in Bixel and Apai (2020b) (hereafter B20), the impact on a population level would be a positive "age-oxygen correlation", wherein older inhabited planets are more likely to have oxygenated atmospheres.

If inhabited planets do tend to evolve towards greater oxygen content over time, then what is the typical timescale for this evolution? Earth underwent major oxygenation events at 2–2.5 Gyr of age and again at ~ 4 Gyr (Lyons et al., 2014), suggesting a ~ 4 Gyr "oxygenation timescale" (Catling et al., 2005). These two events mark the boundaries between the Archean, Proterozoic, and Phanerozoic eras, and correspond to shifts in Earth's redox balance where the amount of oxygen being produced by life became large enough, and/or the geological sinks for oxygen became diluted enough, that oxygen was allowed to build up in the atmosphere. However, a great diversity of planetary factors might affect redox balance, such as outgassing rates, the stellar radiation profile, biogenic
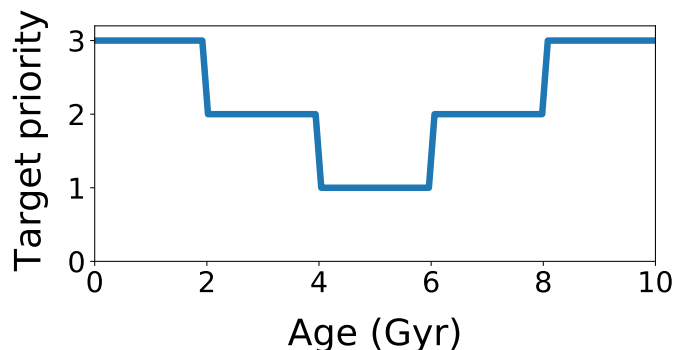
Figure 6.9: Target prioritization for both surveys in Section 6.4.3, optimized to favor observations of younger and older planets to maximize the detectability of age-dependent trends. This also reflects the age distribution of characterized targets, because the simulated planet sample has a uniform age distribution.

oxygen flux, and the planet's initial reducing matter inventory (Catling et al., 2005; Bixel and Apai, 2020b). As a result, Earth's oxygenation timescale could be unusually fast or slow compared to the overall population of inhabited worlds.

A next-generation biosignature survey could not only detect the proposed age-oxygen correlation, but also measure the typical timescale over which this evolution occurs. This measurement could be used to test models for the geological and biological evolution of Earth-like planets and offer insight into how Earth relates to the rest of that population. Here, we assess the ability of direct imaging and transit surveys to study the oxygenation history of Earth-like planets. This section follows a similar methodology to our previous analysis (B20), but expands upon it by incorporating a more thorough assessment of planet occurrence rates, detection sensitivity, and survey strategy, and by studying a broader range of evolutionary timescales.

6.7.1 Model predictions

We assume a fraction f_{life} of EECs to be inhabited by life - note that this parameter absorbs factors affecting both the planet's habitability and the likelihood of life originating. Over time, simulated inhabited planets transition from anoxic to oxygenated atmospheres at an average rate described by a half life $t_{1/2}$. The resulting fraction of habitable planets that

have oxygenated atmospheres as a function of age t_* is:

$$f_{O_2}(t_*) = f_{O_3}(t_*) = f_{\text{life}} \left(1 - 0.5^{t_*/t_{1/2}}\right) \tag{6.15}$$

Note that we assume oxygenated atmospheres to have both O_2 and its photochemical byproduct O_3. We run simulations for f_{life} ranging from 0–100% and for $t_{1/2}$ ranging from 500 Myr – 50 Gyr.

6.7.2 Simulated survey

6.7.2.1 Measurements

The measurements performed by each simulated survey are summarized in Table 6.7. First, we measure the age (t_*) of every planet's host star with 10% precision for the imaging survey and 30% precision for the transit survey. These estimates represent the state of the art for high- and low-mass stars, respectively. For high-mass stars, asteroseismology has yielded highly precise age constraints for *Kepler* targets (e.g., Creevey et al., 2017; Kayhan et al., 2019; Lund et al., 2019), and will likely be able to do so for most of the $\mathcal{O}(100)$ stellar targets probed by an imaging mission. For low-mass stars, asteroseismology has not been successful (e.g., Rodríguez-López et al., 2015; Rodríguez et al., 2016; Berdiñas et al., 2017), and age determination currently relies on a synthesis of model-based estimates. As an example, Burgasser and Mamajek (2017) use a combination of approaches to determine the age of TRAPPIST-1 planetary system (Gillon et al., 2017) with $\sim 30\%$ precision.

Next, each planet is observed to constrain the presence of oxygen. For an Earth-like planet, O_2 can be detected directly through its $0.77\,\mu$m absorption feature or inferred through absorption by stratospheric ozone in the Chappuis (0.40–0.65 μm) or Hartley (0.2–0.3 μm) bands. It should be noted that our calculations assume modern Earth O_2 and O_3 abundances, an assumption which we revisit in Section 6.7.6.1.

For each survey mode, we determine which of these three features would be easiest to observe across the full range of detected EECs. In imaging mode we observe O_2-A absorption; while the Hartley band may be easier to detect for a solar-type star, it

becomes more expensive to observe for lower-mass stars, and the Chappuis band signal is too shallow. Ultimately this consideration is unimportant for the volume-limited imaging survey, and it is likely that all three features will be searched for in the atmospheres of all detected EECs. In transit mode we observe the Chappuis band, as its signal is strong in transit observations. The Hartley band is inaccessible for the vast majority of (predominantly M dwarf) transit survey targets, and the O_2-A feature is too shallow and narrow to detect for distant targets.

In total, the simulated surveys produce measurements of (t_*, O_2) for each observed EEC, where $O_2 = \{0, 1\}$ indicates the detection or non-detection of either O_2-A absorption (imaging mode) or O_3 Chappuis band absorption (transit mode).

6.7.2.2   Target prioritization

Unlike in the previous example, we do not prioritize targets by size or insolation except that we assume all targets have previously been identified as EECs (perhaps with follow-up observations to confirm the presence of H_2O). This assumption is not trivial; imaging surveys cannot easily determine a planet's size, and the true range of planet sizes and insolations that permit habitability are not yet known. In reality, it is likely that an actual biosignature survey will probe some planets that are not habitable for reasons yet unknown to the observer, which will serve as a source of noise (i.e. by reducing f_{life}).

However, we do prioritize targets by age according to Figure 6.9, with observations of the youngest and oldest planets being preferred. This is not intended to counter any bias in the underlying sample, as there are no factors that bias the number of planets that can be characterized by our simulated surveys as a function of age. Rather, as we demonstrate in B20, a survey that prioritizes younger and older planets will be more sensitive to monotonic, age-dependent trends because of the larger contrast between those categories. While this prioritization strategy is optimal for studying the evolution of Earth-like planets, it must be balanced versus the survey's other goals. Notably, it de-prioritizes observations of modern Earth analogs, which may be the best planets to probe if the sole goal is to maximize the chance of detecting O_2.

6.7.2.3 Time budget

As discussed in Section 6.7.2.1, we consider the detection of O_2-A absorption in imaging mode and O_3 Chappuis band absorption in transit mode. The details of the exposure time calculations are shown in Table 6.4. Using PSG, we determine the exposure time required for the reference target to be $t_{ref} = 2.6$ hr for imaging mode and $t_{ref} = 74$ hr for transit mode.

6.7.3 Hypothesis and prior assumptions

Once more, we take the role of an observer intrepreting the results of each simulated survey. Our hypothesis is that inhabited planets tend to evolve towards greater oxygen content over time, and can be stated in similar terms as Equation 6.15:

$$h(t_*) = f_{life} \left(1 - 0.5^{t_*/t_{1/2}}\right) \tag{6.16}$$

We adopt broad, log-uniform prior distributions for f_{life} and $t_{1/2}$, shown in Table 6.3, reflecting our significant prior uncertainty as to frequency and evolutionary timescales of inhabited planets.

6.7.4 Correlation test

In lieu of the Bayesian evidence test used in the previous example, we employ the Mann-Whitney test (Mann and Whitney, 1947) to determine whether t_* correlates with the presence of oxygen, as we previously have done in B20. This model-independent test is more sensitive for detecting the correlation than the Bayesian evidence-based approach, especially in the limit of small sample sizes. However, it does not allow for the estimation of $t_{1/2}$, for which we rely on MCMC sampling.

Example 2: Results for 15-meter imaging survey
How many exo-Earth candidates are probed for O$_2$?

Could this survey detect the age-oxygen correlation?
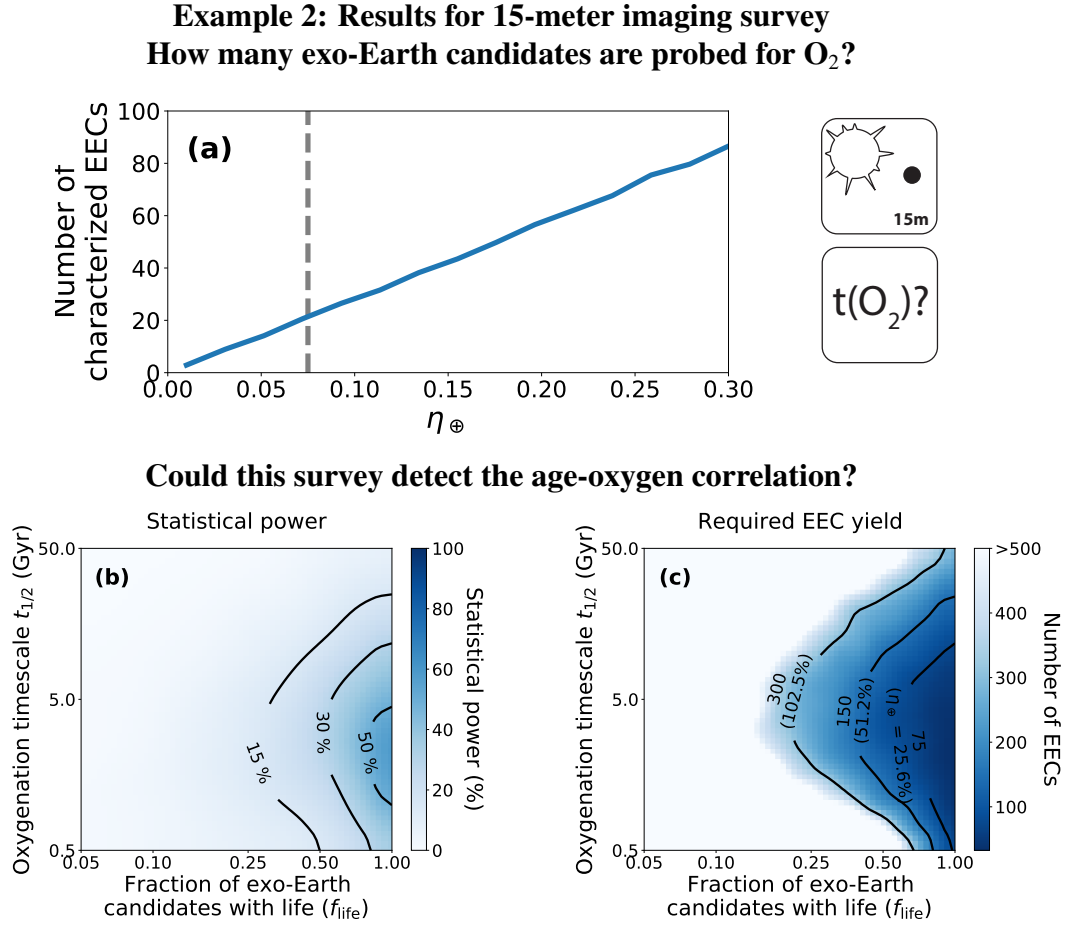
Figure 6.10: Results for the imaging survey in Section 6.7. (a) The number of EECs observed versus $\eta_\oplus$ (for G stars), assuming $t_{\text{total}} = 120$ d. For our baseline case, we set $\eta_\oplus = 7.5\%$. (b) The statistical power to detect the age-oxygen correlation as a function of the astrophysical parameters in Equation 6.15. (c) The minimum number of EECs that must be characterized to achieve 80% statistical power, with the corresponding values of $\eta_\oplus$.

Example 2: Results for 50-meter (equivalent area) transit survey
How many exo-Earth candidates are probed for O_3?
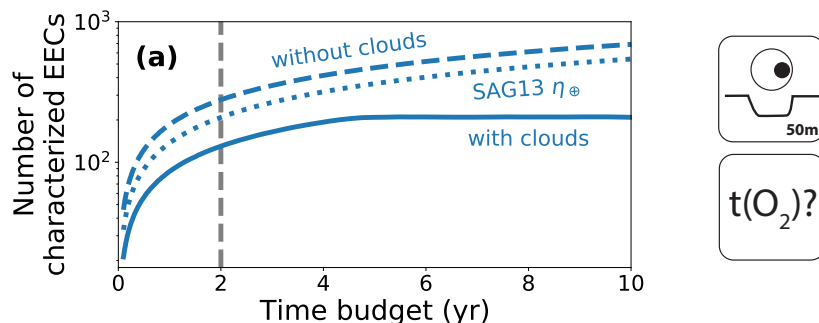
Could this survey detect the age-oxygen correlation?
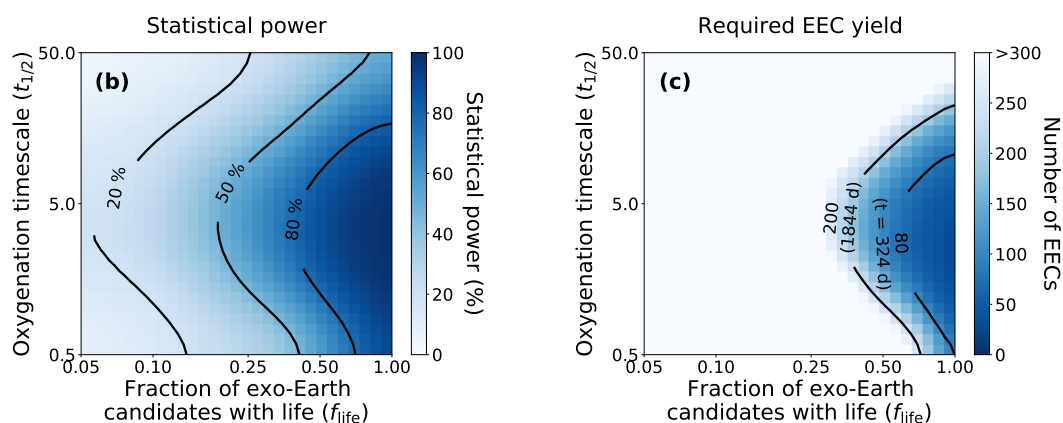
Could this survey determine the oxygenation timescale?
(results for six random realizations of the survey)
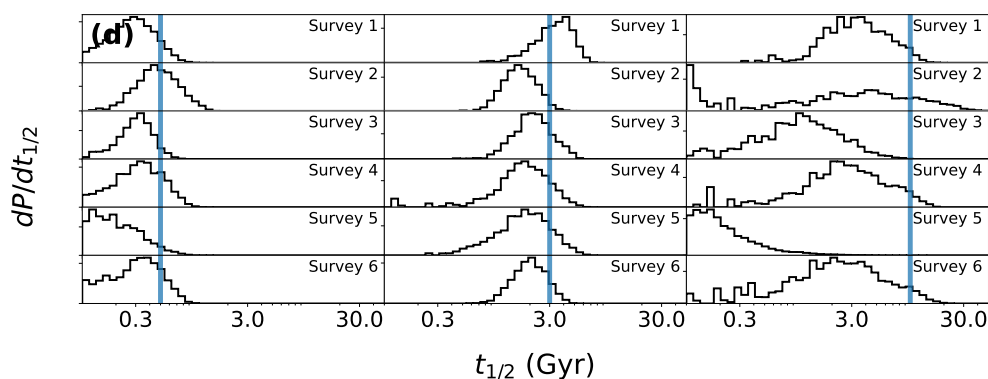
Figure 6.11: Results for the transit survey in Section 6.7. (a) The number of EECs observed versus the observing time budget, assuming $\eta_\oplus = 7.5\%$ for G stars and cloudy atmospheres (solid). 2–4× as many planets could be observed if clouds were neglected (dashed), or 1.5–3× as many with clouds if assuming the higher SAG13 estimate of $\eta_\oplus = 24\%$ (dotted). As our baseline case, we set $t_{total} = 730$ d.

Figure 6.11: (cont.) (b) The statistical power to detect the age-oxygen correlation as a function of the astrophysical parameters in Equation 6.15. (c) The minimum number of EECs that must be characterized to achieve 80% statistical power, with the necessary observing time budget t_{total}. (d) Distribution of possible values for the oxygenation timescale as measured by six random realizations of the survey under the optimistic assumption that 80% of EECs are inhabited. Results are shown for fast (0.5 Gyr, left), Earth-like (3 Gyr, center), and slow (10 Gyr, right) evolutionary scenarios, with the truth values marked by a blue line.

6.7.5 Results

We assess the statistical power of each survey to test the age-oxygen correlation hypothesis, using the Mann-Whitney test to determine whether a positive correlation can be detected in each simulated data set, and `emcee` to sample the posterior distributions of $t_{1/2}$ and f_{life}. Our results are summarized in Figure 6.10 for the imaging survey and Figure 6.11 for the transit survey.

6.7.5.1 Imaging survey

Assuming $\eta_{\oplus} \approx 7.5\%$, it is unlikely (though not impossible) that a direct imaging survey will be able to detect the age-oxygen correlation with a sample of 15–20 EECs. This is generally consistent with our analysis in B20, which suggests a statistical power of 50% for a sample size of ~ 20 EECs only if most are inhabited. In order for an imaging survey to be reliably capable of studying the oxygen evolution of Earth-like planets under optimistic circumstances, a sample size of > 50 EECs is necessary, requiring either $\eta_{\oplus} \gtrsim$ 20% or a smaller inner working angle than assumed here (3.5 λ/D).

6.7.5.2 Transit survey

By probing 100-150 EECs for ozone, the transit survey is able to detect the age-oxygen correlation with high statistical power assuming life to be somewhat common ($f_{\text{life}} \gtrsim$ 50%) and the typical oxygenation timescale to be 1–10 Gyr. If life is very common

($f_{\text{life}} \gtrsim 80\%$), high statistical power can be achieved even if the average oxygenation timescale is as short as ~ 500 Myr or as long as ~ 20 Gyr.

Under the case where life is very common, the transit survey could place meaningful constraints on the oxygenation timescale. As shown in Figure 6.11, the survey can distinguish between scenarios where global oxygenation occurs very quickly ($t_{1/2} \sim 0.5$ Gyr) or at a more Earth-like pace (~ 3 Gyr), but it will be difficult to accurately measure the oxygenation timescale if it is much longer than Earth's ($\gtrsim 10$ Gyr), since no planets of that age exist in the sample. This is due in part to the high degeneracy between $t_{1/2}$ and f_{life} - that is, if only a few oxygenated planets are found, it may be because life is uncommon, or because life is common but global oxygenation is very slow and has not yet had time to occur on most inhabited worlds.

6.7.6 Discussion

6.7.6.1 Detectability of oxygen through Earth's history

In this section we consider all oxygenated planets to have the same O_2 and O_3 abundance as modern Earth. However, during the Proterozoic era (approx. $2.2 - 0.6$ Gya), Earth had a partially oxygenated atmosphere with $pO_2 < 1\%$ (Lyons et al., 2014). If other inhabited planets do evolve like Earth, this suggests that many of them may have 1–3 orders of magnitude less O_2 than modern Earth.

In our analysis, $t_{1/2}$ is the typical timescale require for a planet to achieve a detectable amount of O_2 or O_3. Even if Proterozoic Earth analogs are common and their oxygen is undetectable, our results should not be affected provided that they will eventually develop richly-oxygenated atmospheres like modern Earth's. In this case $t_{1/2}$ corresponds to the end of the Proterozoic (~ 4 Gyr for Earth). On the other hand, it may be that inhabited and oxygenated planets are common but very few of them ever evolve beyond $pO_2 = 0.1 - 1\%$, in which case $t_{1/2}$ corresponds to the end of the Archean (~ 2 Gyr for Earth). In this case, a survey aiming to detect the age-oxygen correlation would need to focus on a smaller number of targets with much deeper observations, and would likely need ultraviolet sensitivity to detect the deep O_3 Hartley band absorption that would have been

detectable throughout the Proterozoic (Reinhard et al., 2017). For transit spectroscopy, ultraviolet sensitivity will be difficult to achieve in a sample of predominantly M stars, so to detect O_3 at Protorezoic-like levels will require the prioritization of G and K targets instead. A LUVOIR-like direct imaging survey targeting G and K dwarfs may be capable of detecting Proterozoic-like ozone levels for individual targets, but the sample size will still be too small unless both $\eta_\oplus$ and f_{life} are large ($\gtrsim 30\%$).

6.7.6.2 Abiotic oxygen sources

We only consider planets on which O_2 is biologically produced - as it was in Earth's history - but others have considered scenarios through which an Earth-sized planet near or within the habitable zone could acquire detectable levels of oxygen through abiotic processes (for a review, see Meadows et al., 2018b). The oxygen in these models tends to initially derive from H_2O or CO_2 dominated atmospheres shortly after the planet's formation and can linger in the atmosphere long enough to serve as a potential "false positive" biosignature for next-generation observatories. In B20, by assuming the fraction of planets with abiotically produced oxygen to be independent of age, we show that these false positives will have a small impact on the detectability of the age-oxygen correlation provided that they are less common than Earth-like planets with biogenic O_2.

In reality, atmospheres with abiotically-produced oxygen will evolve over time. On Earth, oxygen is continually produced in large enough quantities to overcome its substantial geological sinks. On planets where oxygen is, e.g., a remnant of primordial ocean loss, it would be depleted over time. This suggests a statistical test to determine whether oxygen is a reliable biosignature: if the fraction of EECs with oxygen *decreases* with age, this would suggest much of the oxygen to be of a primordial, abiotic origin.

Finally, it is plausible that both populations of oxygen-rich worlds exist in comparable numbers: one with abiotically-produced oxygen that diminishes over time, and another with biologically produced atmospheres that increases over time. Whether the Earth-like age-oxygen correlation could be detected would depend on the timescales of the two processes. For example, if most planets with abiotically produced oxygen lose it before 1 Gyr, and most planets with biogenic oxygen acquire it by 10 Gyr, then it should be

possible to distinguish the two populations.

6.8 Summary

We have presented `Bioverse`, a simulation tool designed to gauge the potential of future observatories to test statistical hypotheses about the formation and evolution of planetary systems and habitable worlds. To achieve this, `Bioverse` leverages statistically realistic simulations of nearby planetary systems, a survey simulator designed to produce data sets representative of different observatory configurations and survey strategies, and a hypothesis testing module to assess the information content of the data. We demonstrated two applications of our code.

In the first example, we determined whether a future direct imaging (15-meter diameter) or transit spectroscopy (50-meter equivalent diameter) survey could empirically test the concept of a habitable zone as well as measure its location and width. With samples as small as 15–20 EECs, we found that both surveys will be capable of testing the habitable zone hypothesis if habitable planets are common ($\gtrsim 50\%$ of EECS), and that they can constrain the habitable zone's width well enough to rule out very wide (e.g., 1–10 AU) or narrow (e.g., 1–1.1 AU). A survey that can characterize 60–70 EECs for atmospheric water vapor can test the habitable zone hypothesis even if habitable planets are less common (20–40% of EECs), but would be difficult to achieve with currently-envisioned direct imaging observatories. Our estimates suggest that this would be feasible for a large aperture transit survey, but the EEC sample size is sensitive to the impact of cloud cover (and other factors not considered here, such as stellar contamination (Rackham et al., 2018)).

In the second example, we expanded upon the age-oxygen correlation proposed in B20, finding that future surveys that aim to study the oxygen evolution of Earth-like planets must expect to characterize at least ~ 50 EECs by detecting the presence of modern Earth-like O_2 or O_3 absorption. With a sample size of 100-150 EECs – if most of them are inhabited – a survey could begin to constrain the evolutionary timescale with meaningful precision, and could determine whether the oxygenation of Earth-like planets proceeds at an Earth-like pace (2–5 Gyr timescale) or much faster (~ 0.5 Gyr). The ability to detect

far-UV O_3 absorption will be beneficial if Proterozoic Earth analogs are common, but may not be necessary provided they eventually evolve to a modern Earth-like state.

The statistical power of either survey to test these hypotheses depends critically on the number of EECs detected, but recent evidence suggests that existing estimates of $\eta_\oplus$ are too high (Pascucci et al., 2019; Neil and Rogers, 2020). Assuming $\eta_\oplus = 7.5\%$ for Sun-like stars, we found that an ambitious 15-meter mirror diameter imaging survey would likely detect 15–20 EECs. Such a survey may have high statistical power for studies of terrestrial planets in general (including those outside the habitable zone), but will only be able to test the habitable zone concept if most EECs are habitable, or if tracers of habitability other than H_2O absorption are considered. Unless $\eta_\oplus > 20\%$, an imaging survey will probably not be able to study the oxygen evolution of truly Earth-like planets, though it might still offer constraints on how common such planets are (Checlair et al., 2021).

In this paper we discussed the statistical power to test hypotheses as a function of sample size given a single measurement for each target. `Bioverse` can also combine multiple measurements for each planet that trace the same underlying physical conditions (such as habitability), allowing surveys to achieve greater statistical sensitivity with limited sample sizes. For example, by incorporating measurements of planetary brightness and color in addition to H_2O absorption, imaging surveys may be able to test the habitable zone concept with smaller sample sizes - provided a hypothesis exists for how these properties should vary with orbital separation (e.g., Checlair et al., 2019). Similarly, if clouds make the detection of H_2O difficult for a transit survey, then stratospheric O_3 may offer an alternative tracer of planetary habitability (provided O_2 is predominantly produced by life).

With `Bioverse`, we aim to enable future space-based exoplanet surveys to test hypotheses including and beyond the examples explored here, and to emphasize the importance of population-level studies for next-generation exoplanet surveys. While target-by-target analyses of the closest planets will be valuable, population-level studies will reveal fundamental truths about the laws governing non-habitable, habitable, and inhabited worlds.

6.8.1 Acknowledgements

CHAPTER 7

Conclusion

7.1 Results

The work I have presented in this thesis focuses on the use of statistics and probabilistic reasoning to promote a better understanding of habitable exoplanets, with a particular emphasis on future efforts to characterize them from space. Following a brief summary of my results in this section, I will conclude with a set of key findings to inform near-future directions for research and technology development relevant to the search for life beyond the solar system.

7.1.1 Bayesian constraints on planetary habitability

In Chapters 2 and 3, I demonstrated how Bayesian methods based on *Kepler* statistics can be used to transform limited information about individual planets into robust probabilistic constraints on their potential habitability. In Chapter 2, we explored the likely composition of Proxima Centauri b. Folding its host spectral type and minimum mass limit into a Bayesian analysis based on exoplanet occurrence rates and mass-radius relationships, we found that it was $\sim 90\%$ likely to be a terrestrial world. A similar analysis could be applied for future radial velocity detections of planets as a means of prioritizing targets for ground-based ELTs. Meanwhile, future investigations into the relationship between planet size and composition will further our understanding of Proxima Cen b and its analogs.

In Chapter 3, I developed and assessed a Monte Carlo approach for identifying potentially habitable planets in direct imaging data. This analysis offers a solution to a known problem for future LUVOIR- or HabEx-like missions - namely, that any potentially habitable planets detected by such a telescope will be difficult to distinguish from the many more numerous "false positives" (Guimond and Cowan, 2018). If correctly applied, this

method could spare weeks of observing time on a flagship-class mission. Furthermore, our assessment of the relative information content offered by different types of measurements should inform instrumentation choices. For example, while mass measurements are useful for classifying planets, they are far from definitive due to the observed variability in exoplanet densities. On the other hand, multi-band photometry can efficiently select planets, but only if the relative colors of different types of planets can be accurately predicted.

7.1.2 Probabilistic solutions to stellar contamination in transit spectra

In Chapter 4, I assessed the transit spectrum of the hot Jupiter WASP-4b. Our analysis presented a featureless spectrum, consistent with previous studies of this target in the literature, and featured a more thorough treatment of stellar activity than most published transit spectra to date. In particular, we used Bayesian model selection techniques to determine whether sufficient evidence existed for stellar contamination in the transit spectra, and also used our data to assess the size and temperature of an active region occulted by the planet during one of our observations. This type of analysis demonstrates a path forward for piecing apart the stellar and planetary components of transmission spectra, which will be vital for future efforts probe the atmospheres of transiting habitable planets.

7.1.3 Statistical hypotheses for next-generation exoplanet observatories

In Chapters 5 and 6, I presented and developed hypotheses for statistical relationships between the properties of terrestrial exoplanets and explored how they could be tested with next-generation space-based observatories. In Chapter 5, I forwarded a new hypothesis for the atmospheric evolution of habitable planets based on our understanding of Earth's history. Namely, if other habitable worlds are truly Earth-like, they might evolve towards greater oxygen over time, leading to a positive "age-oxygen correlation". This hypothesis could be tested by next-generation observatories, and if such a correlation is found, it will confirm that Earth's oxygen evolution is typical for a habitable world. Such a discovery would also strongly suggest that oxygen is a reliable biosignature for other planets, as it

is for Earth.

Finally, in Chapter 6, I presented a general framework for evaluating the statistical potential of next-generation space missions, called `Bioverse`. `Bioverse` evaluates both direct imaging and transit spectroscopy surveys of terrestrial planets, and incorporates estimates of the sample sizes and statistical biases inherent to each technique. We demonstrated that future missions will be capable of detecting the presence of a habitable zone by examining how H_2O abundance as a function of insolation, but only if habitable planets are relatively common. Furthermore, a transit survey that searches for O_3 in the atmospheres of potentially habitable worlds may be capable not only of detecting the "age-oxygen correlation" proposed in Chapter 5, but also measuring the timescale over which habitable planets become oxygenated. `Bioverse` is generally applicable to scenarios beyond the two considered here, and could play a valuable role in developing the science case of next-generation habitable exoplanet observatories.

7.2 Key findings

Toward the end of Chapter 1, I laid out important questions regarding future efforts to characterize habitable exoplanets. Here, I will summarize my findings on how to best answer these questions from a statistical point of view.

7.2.1 Population studies should play a crucial role in future space missions

Kepler has arguably had a greater impact on exoplanet science than any other mission to date. This was due not to what it revealed about any individual exoplanet, but what it revealed through its thousands of detections about the exoplanet population as a whole. By the same token, I argue that statistical studies of terrestrial exoplanets will offer greater insight into the fundamental questions in astrobiology than even the most detailed characterizations of individual worlds. It is therefore critical that next-generation observatories be evaluated on the basis of their ability to conduct these studies. This ability depends not only on the overall predicted sample size, but the technique by which the planets are observed and the resources required to properly characterize them. `Bioverse`, which

I presented in Chapter 6, provides a rigorous and flexible framework for evaluating missions on this basis.

The development of specific and testable hypotheses is a crucial step in the scientific method and thus in the effort to study the population of habitable exoplanets. It is therefore necessary that considerable effort be directed toward developing hypotheses about how the properties of habitable exoplanets will vary over the range of parameters they are expected to possess. For example, the hypothesis about the atmospheric evolution of Earth-like planets that I proposed in Chapter 5 could be expanded with more detailed models for redox evolution over a broader range of planetary and stellar parameters. Encouragingly, a growing body of literature is concerned with forging testable predictions for how the measurable properties of habitable planets will vary across the habitable zone (e.g., Bean et al., 2017; Checlair et al., 2019; Lustig-Yaeger et al., 2019b; Turbet et al., 2019; Lehmer et al., 2020). While most of these hypotheses will not be testable for more than a decade yet, they are needed to inform fundamental spacecraft design choices that must be cemented several years prior to launch. Additionally, hypotheses developed before a systematic survey of habitable planets is conducted will be more robust to confirmation bias than those developed after the data already exist, and such safeguards are crucial when contemplating such a groundbreaking potential discovery as the detection of life beyond Earth.

7.2.2   Exoplanet statistics can complement targeted analyses

Just as the combined observations of several individual planets can offer insight into the broader exoplanet population, population-level knowledge can be used to constrain the properties of individual planets through Bayesian inference. Such constraints should play an important role in planning follow-up observations. For example, my analysis in Chapter 2 demonstrates that Proxima Cen b is quite likely to be a terrestrial planet, bolstering the case for resource-intensive efforts to directly image it or even send spacecraft to its system[1]. Similar analyses should be standard for all newly-discovered planets detected through the radial velocity method to better understand their likelihood of habitability.

---

[1]For example, see the Breakthrough Starshot concept here.

Similarly, leveraging *Kepler* statistics to optimize the target prioritization strategy of a direct imaging mission, as demonstrated in Chapter 3, could save weeks of follow-up observing time (and in the case of a starshade-equipped telescope, significant amounts of fuel).

Furthermore, even in-depth observations of individual planets may be difficult to interpret without statistical context. For example, while Earth's O_2 is entirely produced by life, abiotic O_2 sources may play a substantial role in the atmospheres of terrestrial exoplanets, casting doubt on its usefulness as a biosignature. Contextual information - such as the abundances of other species in the atmosphere - is key to understanding whether O_2 is indicative of life (Meadows et al., 2018b). As I argue in Chapter 5, determining how oxygen evolves over time could provide an important piece to this puzzle: by observing a large enough sample of habitable planets, one could determine whether their oxygen content evolves similarly to Earth - in which case it is likely to be biogenic as well - or differently. Conversely, it may be the case that without statistical context, it will be impossible to definitively detect the presence of life beyond the solar system.

Similarly, whether or not atmospheric H_2O is a reliable indicator of the presence of oceans on a terrestrial world remains undetermined. The habitable zone hypothesis predicts that planets with oceans should be more common within a finite range of planet-star distances, and it seems likely that such planets would have H_2O in their atmospheres similar to Earth. On the other hand, a planet that has undergone a runaway greenhouse effect will evolve through a phase where its atmosphere is rich in water vapor despite having no oceans or potential for life - but most such planets would likely be found at higher insolation than Earth. In Chapter 6, I show that by investigating how H_2O abundance correlates with planet-star distance, we can validate the habitable zone hypothesis and prove that the presence of H_2O is a reliable predictor of the presence of oceans.

Our understanding of individual planets' atmospheres will always be limited when observing them from interstellar distances. Statistical context provided by observations of dozens to hundreds of other worlds will allow us to more confidently determine which planets are habitable and which planets host life - and without this context, such determinations may be difficult or impossible.

7.2.3 Solar system science will enable testable statistical hypotheses for future exoplanet surveys

Thus far, exploring the solar system has provided the best basis for developing testable predictions about exoplanet habitability. State-of-the-art models for the habitable zone depend on a complex understanding of Earth's carbon cycle first developed through geological research (Kasting et al., 1993). The study of the atmospheres of Mars and other solar system bodies has motivated more general predictions for the relationship between planetary size and habitability that could be tested in the future (Zahnle and Catling, 2017). Further exploration within the solar system will allow us to better refine these hypotheses and to develop new ones. For example, a more complete understanding of what factors set the pace of Earth's oxygenation will allow for more detailed predictions of the atmospheric evolution of Earth-like planets. It has been suggested that Venus was habitable as recently as ~ 1 Gya (Way et al., 2016) - if this is verified by new Venus orbiters and landers, we might predict that exoplanets analogous to Venus will undergo an extended habitable period before becoming desiccated. This prediction could be tested through observations of Venus analogs across a wide range of ages. Finally, the detection anywhere in the solar system of life with no common ancestry to Earth's would suggest *a priori* that inhabited exoplanets are quite common, and that potential biosignatures such as O_2 are even more likely to have been produced by life than previously thought.

7.2.4 Direct imaging versus transit spectroscopy for statistical exoplanet science

In general, current concepts for next-generation space telescopes designed to study habitable exoplanets focus on either direct imaging or transit spectroscopy as their primary means of atmospheric characterization. The work I presented in Chapter 6 focused on the strength of either technique for conducting statistical characterizations of the terrestrial exoplanet population, but the resources may only exist to fully pursue one of these techniques. There is as yet insufficient research to conclusively determine which technique offers a better path toward this goal, but my work has defined key outstanding questions that must be resolved in order to make this determination.

Direct imaging offers a number of unique benefits, such as the ability to characterize the closest exoplanets to Earth (most of which do not transit their stars) or to study seasonal and rotational variability. Perhaps most importantly, space-based direct imaging is optimized for the study of Earth analogs whose stars are similar to our Sun. This advantage may prove crucial given recent concerns about the habitability of planets around low-mass stars due to these stars' tumultous pre-main sequence evolution (Barnes et al., 2016) and enhanced flare activity (Howard et al., 2018). However, to conduct statistical studies of habitable planets would require the characterization of at least $\mathcal{O}(10)$ exo-Earth candidates, and the recent uncertainty as to the true value of $\eta_\oplus$ raised by Pascucci et al. (2019) casts doubt on whether even the ambitious 15-meter LUVOIR concept could accomplish this goal. This obstacle could be overcome with an even larger mirror or more precisely designed coronagraphic imager, both of which could decrease the IWA but are also likely to significantly inflate development costs. The true value of $\eta_\oplus$ is therefore the primary source of uncertainty in determining the potential of space-based direct imaging for statistical studies, and resolving this uncertainty should be prioritized as a short-term goal.

The exoplanet detection yield of a transit spectroscopy survey will also depend on $\eta_\oplus$ - however, the evidence shows that $\eta_\oplus$ is higher by a factor of $\sim 2 - 4\times$ for low-mass stars as for the typical *Kepler* target (Mulders et al., 2015b). Furthermore, this yield is technologically determined by total light-collecting area - *not* aperture diameter. The *Nautilus* concept exploits this fact by relying a large array of reproducible telescopes whose cost could be expected to scale approximately linearly with light-collecting area. These factors suggest transit spectroscopy to be a promising option for statistical studies even if $\eta_\oplus$ is small (perhaps 5–10%, based on Chapter 6), but this depends on resolving two outstanding astrophysical obstacles. First, an increasing body of research has focused on the impact of stellar heterogeneity on transmission spectra, and its potential to induce false signals in terrestrial planet transit spectra has already been demonstrated from space (Zhang et al., 2018). Possible solutions include using precise stellar spectra to model the heterogeneous composition of the photosphere (Wakeford et al., 2019), conducting

simultaneous visible and infrared transit observations from space[2], and jointly modeling the stellar and planetary features as demonstrated in Chapter 4. The second obstacle is the predicted impact of high-altitude clouds on absorption features due to lower-altitude species, particularly for tidally-locked planets in the habitable zones of M dwarfs. A recent analysis by Pidhorodetska et al. (2020) suggests that even the 15-meter LUVOIR would be incapable of characterizing exo-Earth candidates in the TRAPPIST-1 planetary system - despite its relative proximity to the Sun - due to predictions for severe cloud cover on their daysides. Future research should focus on whether this cloud cover is consistent from transit to transit, its dependence on planetary and stellar parameters, and the vertical distribution of key species such as O_2 and O_3 relative to the clouds. In particular, if most O_3 exists above the clouds - as is the case on Earth - then it should be detectable even with complete cloud cover. Observational tests to deduce cloud cover conditions for planets orbiting the nearest low-mass stars should be prioritized, both through direct imaging with ELTs and transit observations with space telescopes such as JWST or the *Nautilus* Probe-class mission (Apai et al., 2019).

Finally, while it is natural to view direct imaging and transit spectroscopy as competing priorities in the design of next-generation flagship space observatories, it is also important to highlight their complementarity, as the planetary parameters and stellar mass regimes they probe are often non-overlapping. Direct imaging favors planets around solar-mass stars and can be used to study surface features and seasonal variability, while transit spectroscopy favors planets around low-mass stars and can directly constrain their sizes. To study how habitability depends on stellar mass will likely require the combination of both techniques, and combining the unique parameters that each can probe will provide statistical insights that neither is individually capable of.

7.3 Summary

The nearest potentially habitable exoplanet, Proxima Centauri b, is over four light-years away. The most ambitious concepts for its robotic exploration, if successful, would yield

---

[2]See the Pandora SmallSat mission concept.

only preliminary in situ data before the end of this century, and would be inapplicable to most of the hundreds of potentially habitable worlds in the solar neighborhood. Even with advanced telescopes and instrumentation, we will never attain as complete an understanding of any terrestrial exoplanet as we have already acquired for Venus, Earth, and Mars. Instead, the key advantage of exoplanets is that their number and diversity offer statistical insights that the solar system cannot, and it is critical that future missions intended to search for habitable worlds are designed with this advantage in mind.

To this end, I have demonstrated multiple examples of statistical strategies for enhancing our understanding of habitable exoplanets, both now and in the future using next-generation space telescopes. By leveraging these strategies, exoplanet researchers can begin to uncover the laws governing habitability and life in the Universe, and to place Earth into context as one member of the broader population of potentially habitable worlds.

APPENDIX A

Supplementary Materials for Chapter 4

A.1 Binned light curves and posterior distributions

In Figures A.1-A.4 we present the binned light curves for Transits 1–4, along with their fitted models and residuals. We plot the joint posterior distributions of the parameters of the white light curve models in Figures A.5-A.8, and the same for a selection of the retrieval models in Figure A.9.
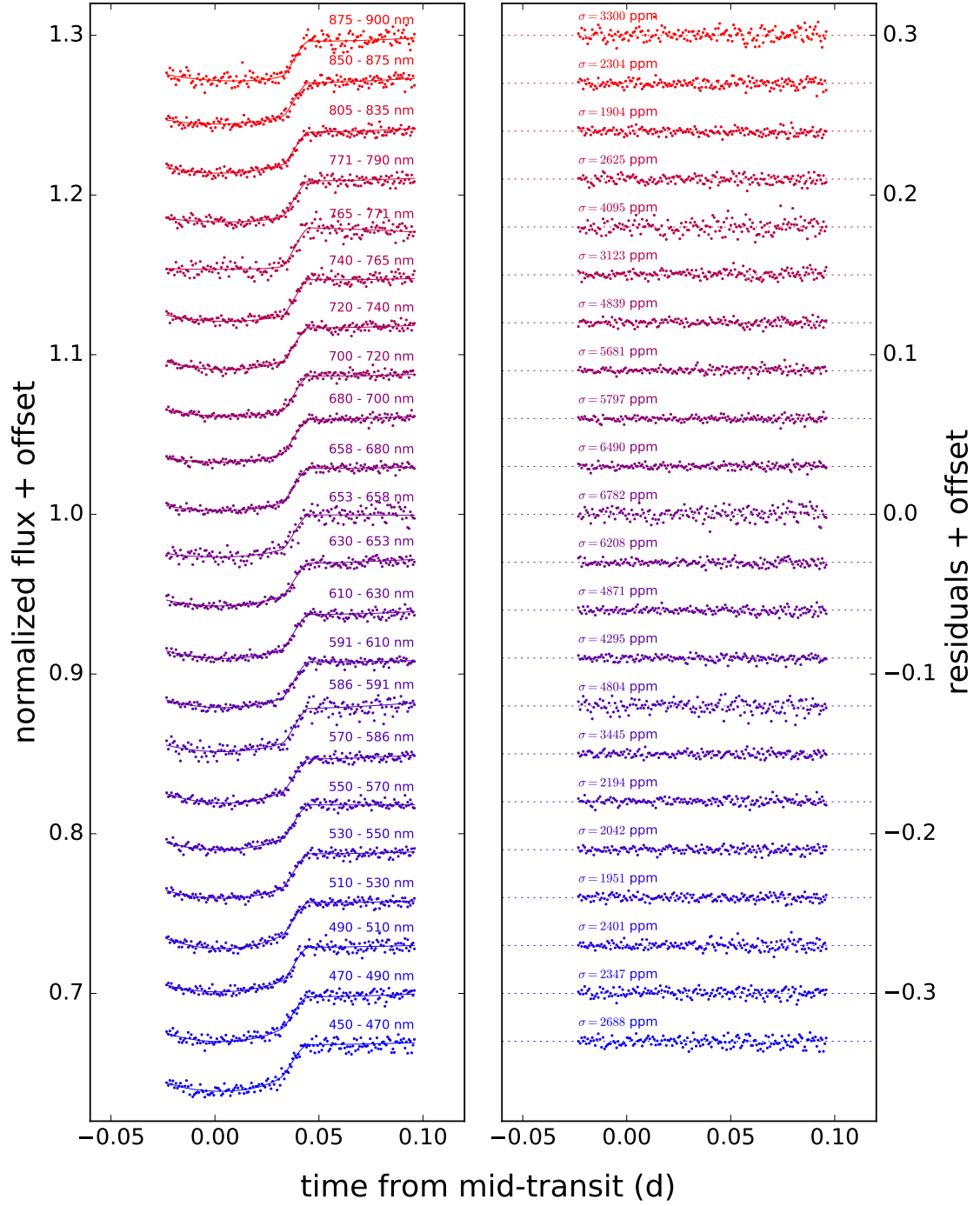
Figure A.1: Binned light curve fits from the night of 2013-09-24. The models in the left panel include the linear systematics component.
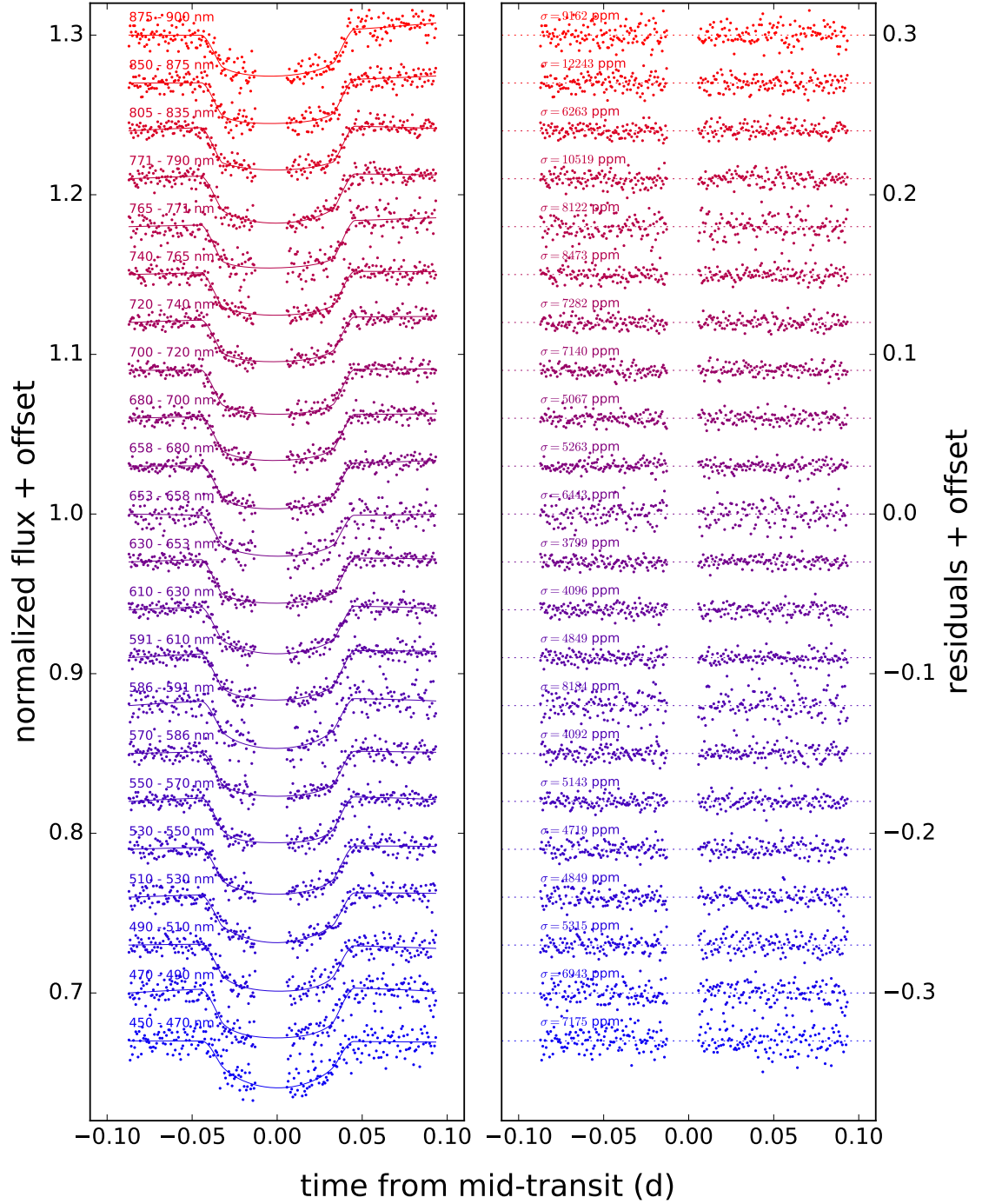
Figure A.2: Binned light curve fits from the night of 2013-10-17. The models in the left panel include the linear systematics component. A potential spot-crossing event is excluded from the fit.
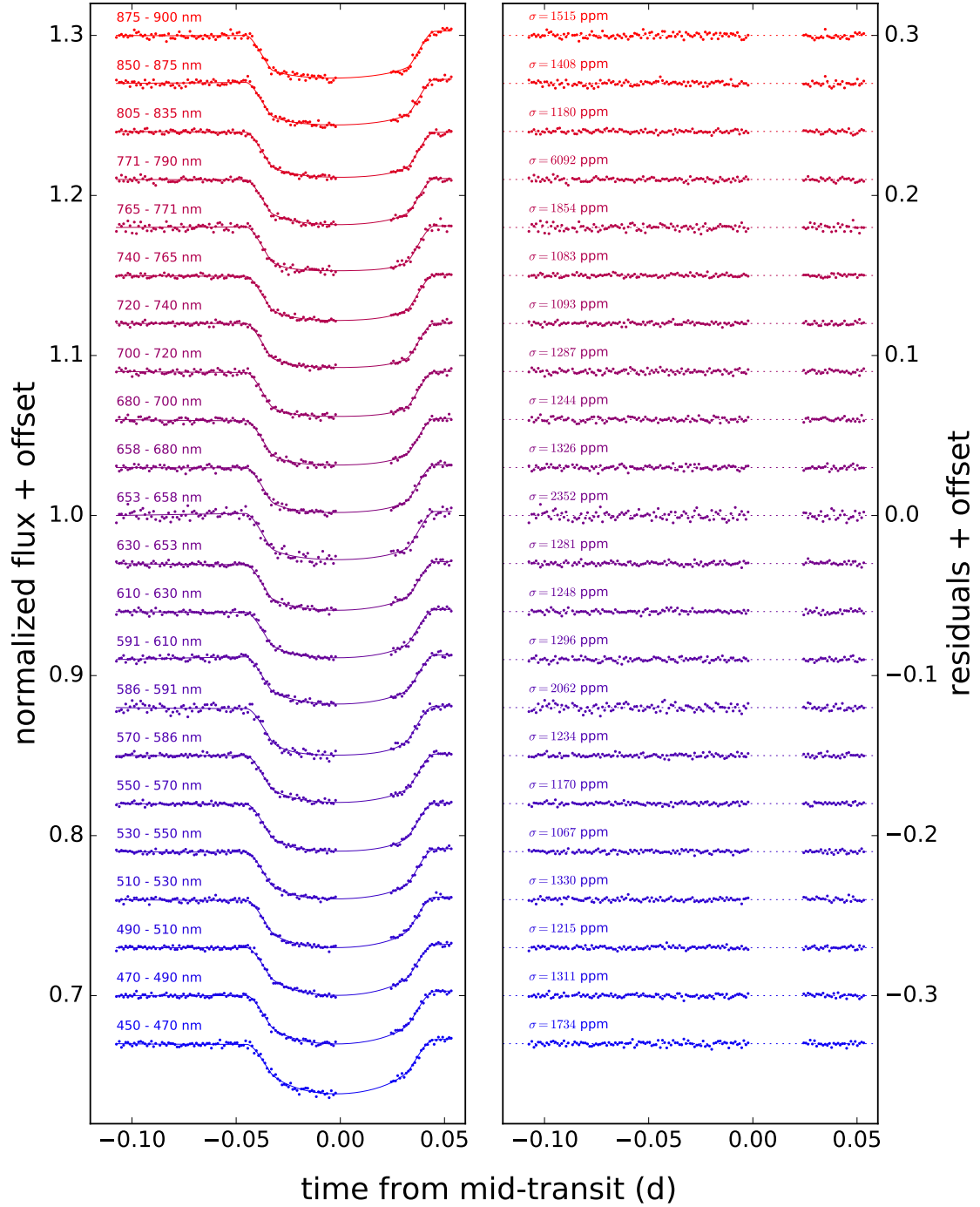
Figure A.3: Binned light curve fits from the night of 2015-08-14. The models in the left panel include the linear systematics component. The spot-crossing event is excluded.
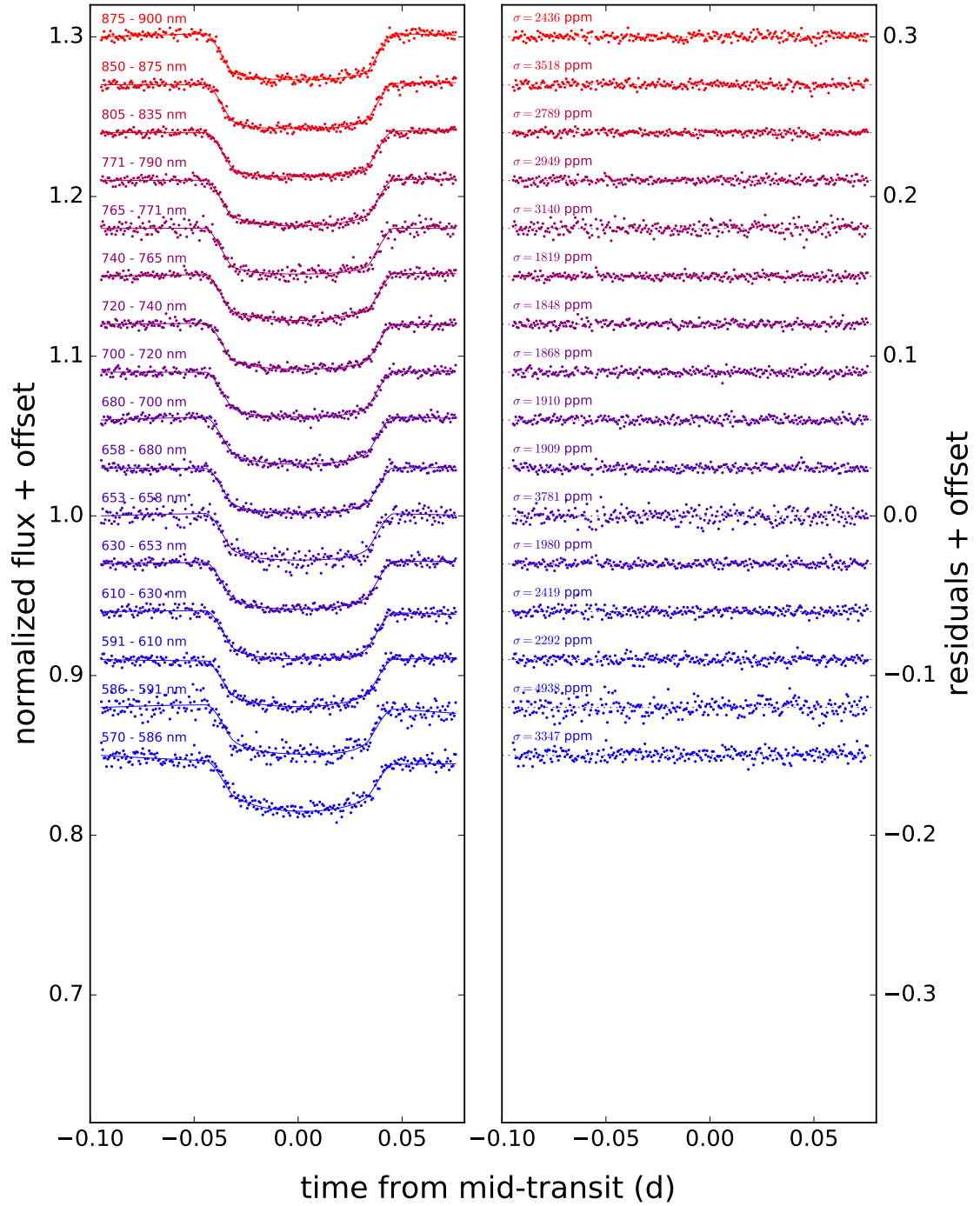
Figure A.4: Binned light curve fits from the night of 2015-09-26. The models in the left panel include the linear systematics component. Six of the bins are excluded for lack of data due to the narrower filter.
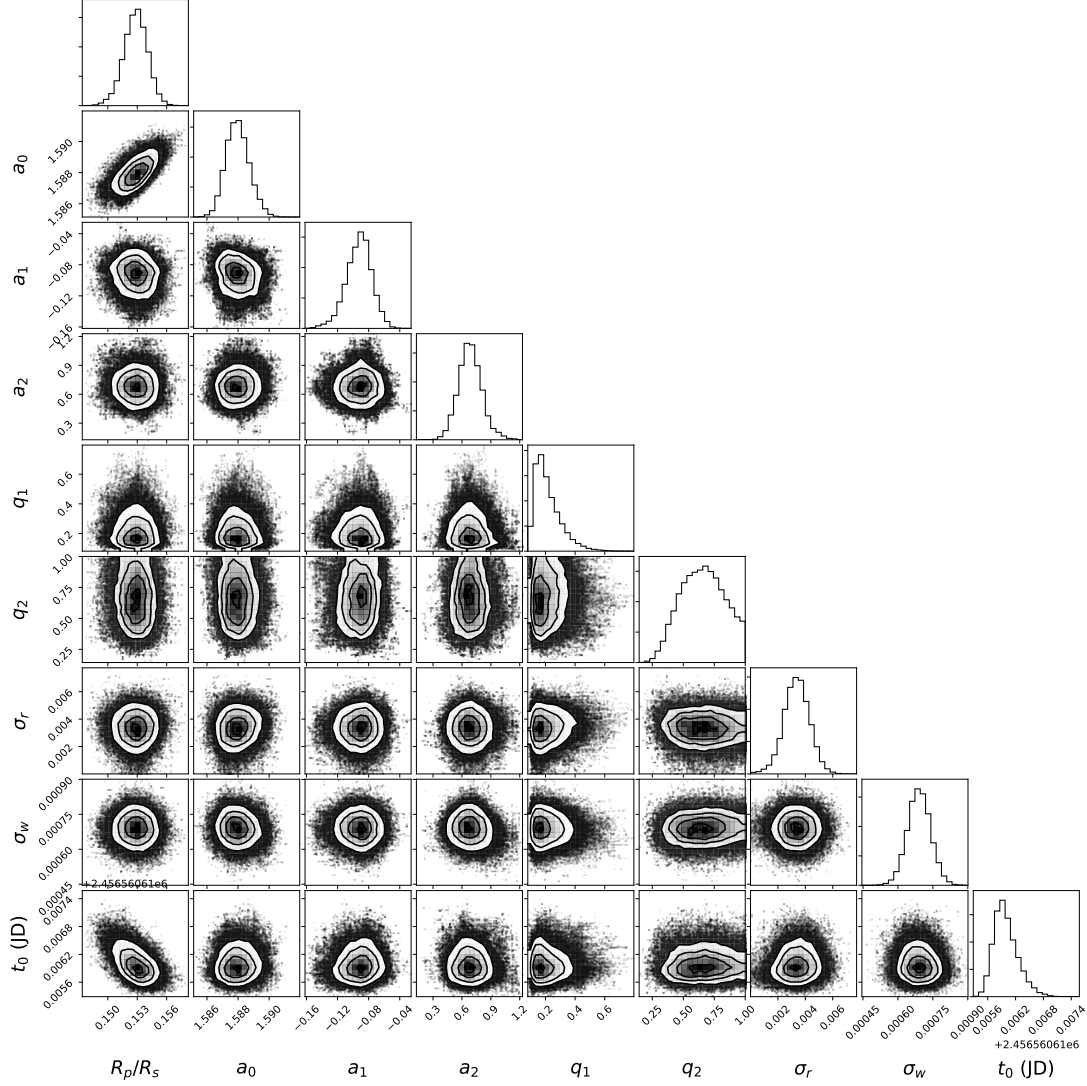
Figure A.5: Joint posterior distributions for the white light curve model for Transit 1, including the planet-to-star radius ratio (R_p/R_s), the correlated and uncorrelated noise parameters (σ_r, σ_w), the mid-transit time (t_0), three parameters for the quadratic polynomial systematics model (a_0, a_1, a_2), and two parameters for the quadratic limb-darkening law (q_1, q_2). The light curve of this transit is incomplete, so we exclude its transit spectrum from our combined result.
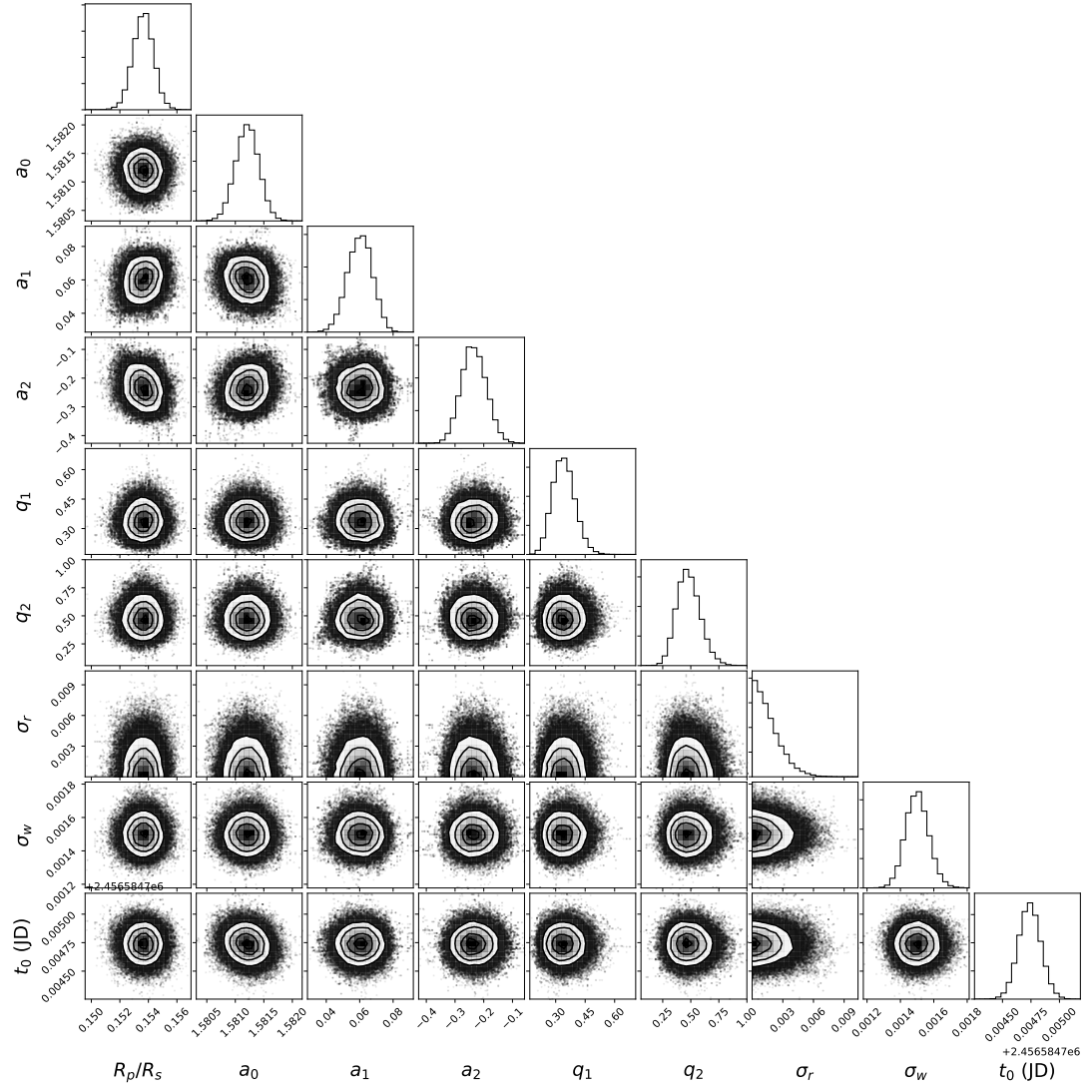
Figure A.6: Joint posterior distributions for the white light curve model for Transit 2.
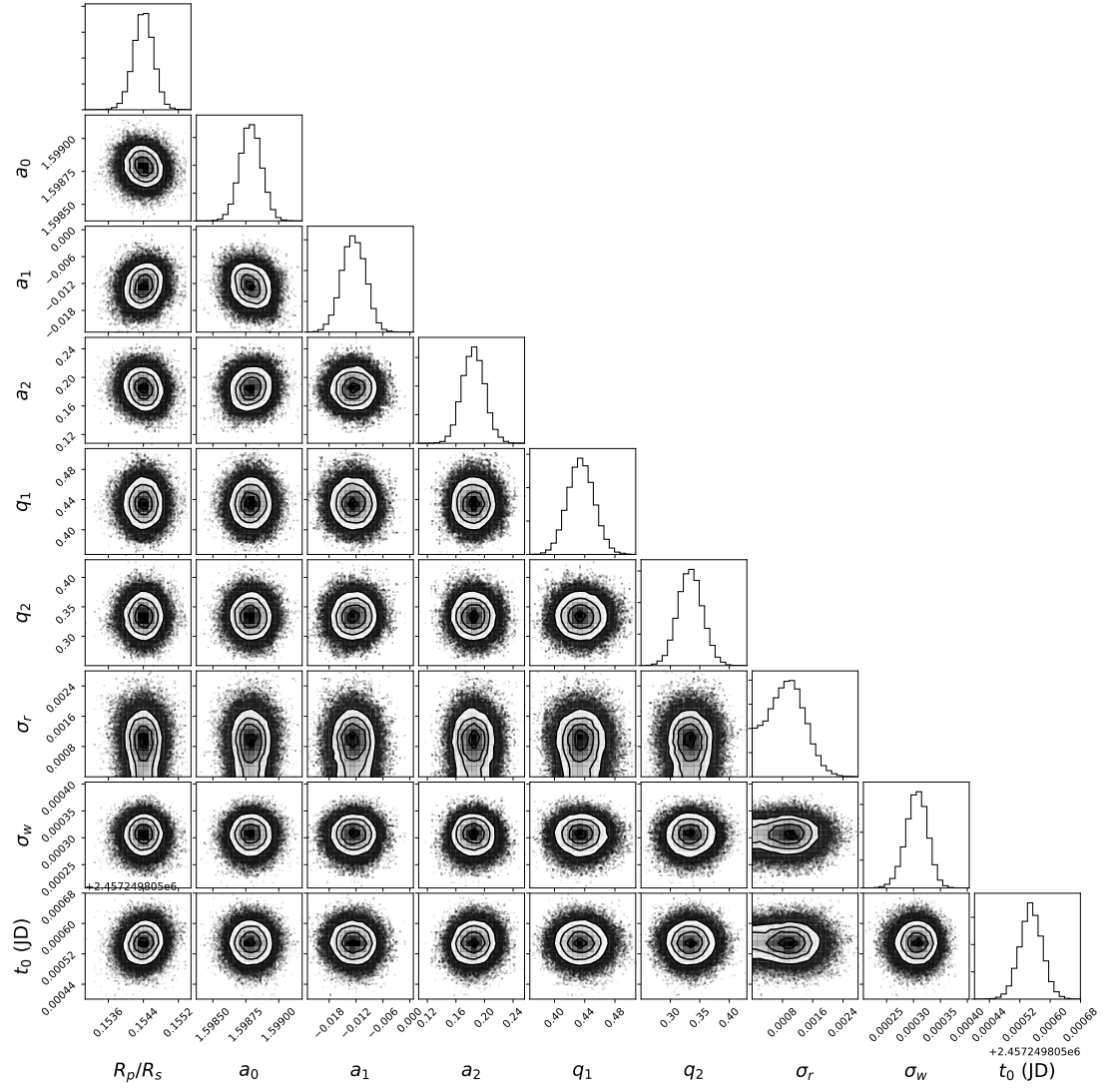
Figure A.7: Joint posterior distributions for the white light curve model for Transit 3.
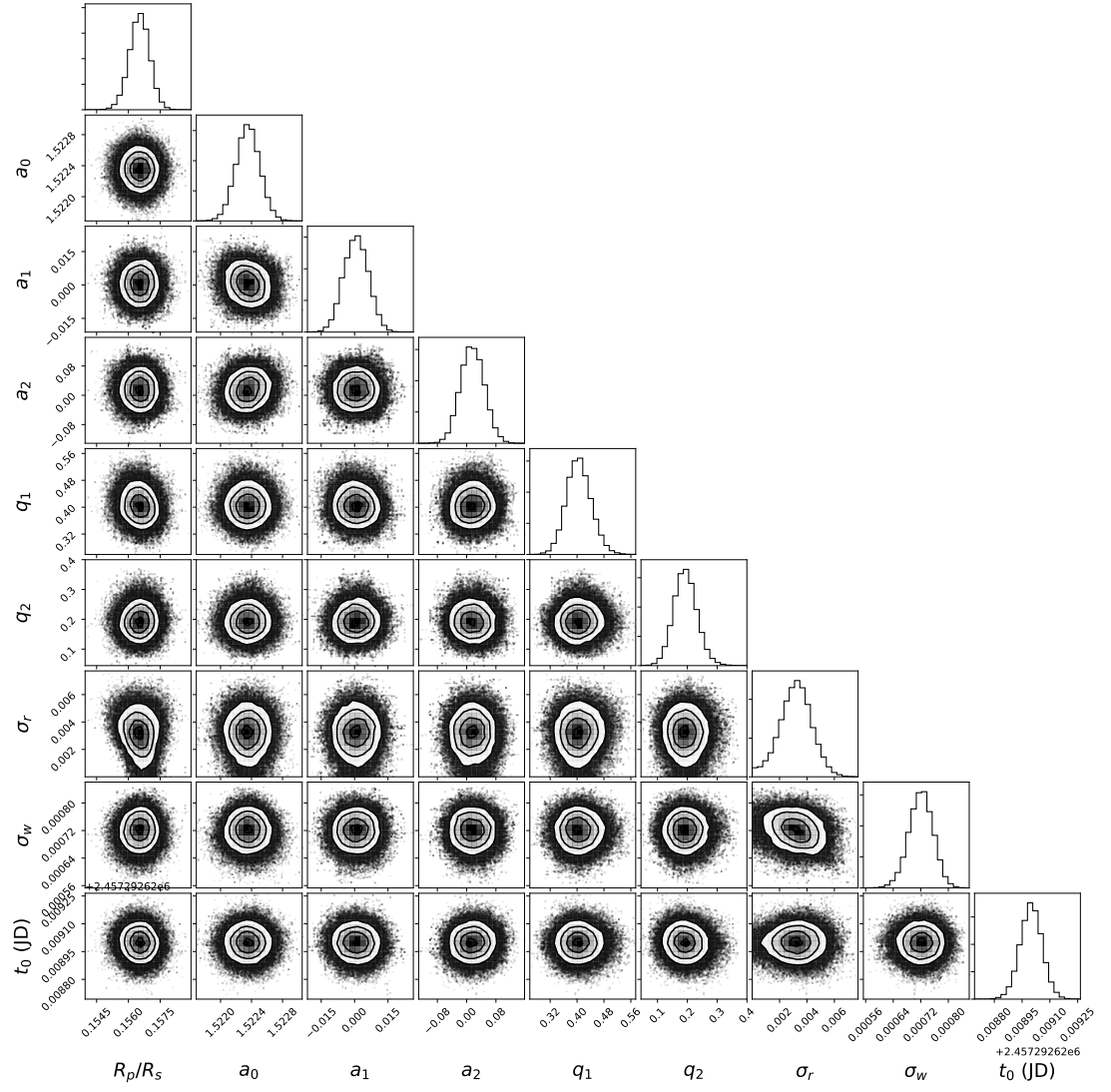
Figure A.8: Joint posterior distributions for the white light curve model for Transit 4.
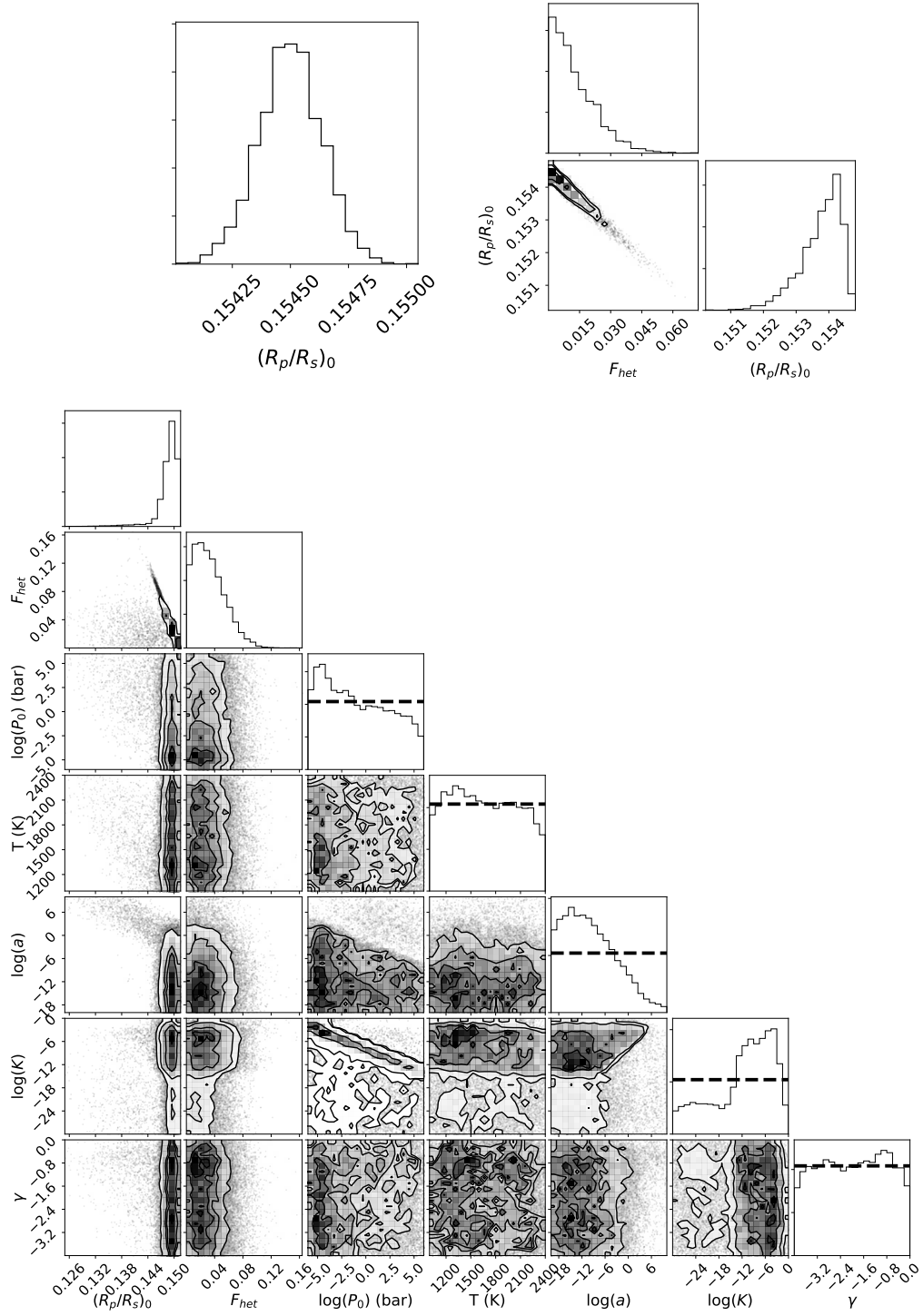
Figure A.9: Joint posterior distributions for a subset of the atmospheric retrieval models shown in Figure 4.8. The parameters and prior distributions are described in Table 4.5.

Figure A.9: (cont.) (Top) Models for a uniform opacity atmosphere with (right) and without (left) the contamination features from the photosphere. Here the posterior distributions have converged, and the degeneracy between the planet's radius and the spot covering fraction is apparent. (Bottom) Model for a hazy atmosphere with K absorption and contamination from the photosphere. Some of the parameters do not converge enough from their priors to yield meaningful constraints. In these cases, the prior distributions are marked with a dashed line.

A.2 Tabulated transmission spectra

In Table A.1 we present the tabulated transmission spectra as well as the correction derived for the effect of the occulted spot of Transit 3.

Table A.1: Data for the combined and individual transmission spectra shown in Figures 4.3 and 4.4 with 1σ uncertainties. The second column is the weighted mean of the mean-subtracted values from Transits 2-4, where the Transit 3 values have first been corrected for the presence of an occulted spot. The last column is the multiplicative effect on the fitted radii for Transit 3 due to the occulted spot, where the true radius $R_{p,0} = R_p/\epsilon^{1/2}$.

Bin (nm)	$\Delta R_p/R_s$	Transit 1	Transit 2	Transit 3	Transit 4	$\epsilon^{1/2}$
450.0 - 470.0	-0.0015 ± 0.0008	0.1525 ± 0.0031	0.1430 ± 0.0028[‡]	0.1531 ± 0.0008		1.0121
470.0 - 490.0	0.0004 ± 0.0007	0.1508 ± 0.0026	0.1583 ± 0.0045	0.1548 ± 0.0007		1.0113
490.0 - 510.0	-0.0015 ± 0.0007	0.1528 ± 0.0024	0.1514 ± 0.0033	0.1531 ± 0.0007		1.0116
510.0 - 530.0	-0.0007 ± 0.0008	0.1543 ± 0.0020	0.1533 ± 0.0029	0.1538 ± 0.0008		1.0118
530.0 - 550.0	0.0007 ± 0.0006	0.1547 ± 0.0018	0.1550 ± 0.0029	0.1551 ± 0.0006		1.0105
550.0 - 570.0	0.0001 ± 0.0005	0.1549 ± 0.0019	0.1530 ± 0.0021	0.1544 ± 0.0005		1.0097
570.0 - 586.8	0.0002 ± 0.0006	0.1528 ± 0.0018	0.1520 ± 0.0025	0.1551 ± 0.0007	0.1543 ± 0.0016	1.0092
586.8 - 591.8[†]	-0.0000 ± 0.0011	0.1529 ± 0.0038	0.1532 ± 0.0046	0.1543 ± 0.0014	0.1569 ± 0.0020	1.0096
591.8 - 610.0	0.0015 ± 0.0006	0.1582 ± 0.0020	0.1574 ± 0.0021	0.1566 ± 0.0007	0.1545 ± 0.0013	1.0089
610.0 - 630.0	0.0002 ± 0.0006	0.1550 ± 0.0019	0.1552 ± 0.0023	0.1533 ± 0.0008	0.1591 ± 0.0011	1.0089
630.0 - 653.8	0.0007 ± 0.0005	0.1492 ± 0.0022	0.1547 ± 0.0020	0.1546 ± 0.0007	0.1579 ± 0.0010	1.0085
653.8 - 658.8[†]	0.0010 ± 0.0012	0.1524 ± 0.0040	0.1442 ± 0.0033	0.1576 ± 0.0017	0.1581 ± 0.0019	1.0076
658.8 - 680.0	-0.0008 ± 0.0006	0.1523 ± 0.0015	0.1496 ± 0.0022	0.1549 ± 0.0008	0.1546 ± 0.0010	1.0082
680.0 - 700.0	0.0005 ± 0.0006	0.1516 ± 0.0018	0.1543 ± 0.0022	0.1535 ± 0.0008	0.1585 ± 0.0008	1.0079
700.0 - 720.0	-0.0005 ± 0.0006	0.1537 ± 0.0019	0.1546 ± 0.0021	0.1539 ± 0.0008	0.1556 ± 0.0009	1.0078

Bin (nm)	$\Delta R_p/R_s$	Transit 1	Transit 2	Transit 3	Transit 4	$\epsilon^{1/2}$
720.0 - 740.0	-0.0002 ± 0.0005	0.1522 ± 0.0022	0.1514 ± 0.0024	0.1537 ± 0.0006	0.1572 ± 0.0008	1.0076
740.0 - 765.0	0.0002 ± 0.0005	0.1559 ± 0.0029	0.1526 ± 0.0028	0.1525 ± 0.0007	0.1588 ± 0.0007	1.0074
765.0 - 771.0†	0.0017 ± 0.0009	0.1591 ± 0.0049	0.1577 ± 0.0045	0.1548 ± 0.0011	0.1601 ± 0.0016	1.0073
771.0 - 790.0	0.0001 ± 0.0006	0.1529 ± 0.0026	0.1571 ± 0.0031	0.1536 ± 0.0008	0.1573 ± 0.0009	1.0071
805.0 - 835.0	-0.0004 ± 0.0006	0.1475 ± 0.0020	0.1553 ± 0.0026	0.1539 ± 0.0007	0.1555 ± 0.0010	1.0070
850.0 - 875.0	-0.0014 ± 0.0007	0.1482 ± 0.0026	0.1503 ± 0.0032	0.1527 ± 0.0011	0.1556 ± 0.0010	1.0066
875.0 - 900.0	-0.0001 ± 0.0010	0.1530 ± 0.0040	0.1495 ± 0.0040	0.1539 ± 0.0010	0.1608 ± 0.0011‡	1.0064

† Denotes bins centered on Na, Hα, and K absorption lines.

‡ These outliers are excluded from the analysis; see Section 4.4.2.

APPENDIX B

Supplementary Materials for Chapter 6

B.1   List of symbols

Table B.1 provides definitions for abbreviations and symbols used in this paper.

Table B.1: A list of common abbreviations and symbols used in this paper.

Symbol	Description
Abbreviations	
EEC	"exo-Earth candidate" (or "potentially habitable planet"); planets in the radius range $0.8(S/S_{\oplus})^{0.25} < R_p < 1.4 R_{\oplus}$
LUVOIR	Large UV/Optical/Infrared Surveyor (The LUVOIR Team, 2019)
SAG13	NASA's Exoplanet Program Analysis Group Science Analysis Group 13
PSG	NASA/GSFC Planetary Spectrum Generator (Villanueva et al., 2018)
IWA, OWA	Inner, outer working angles of a coronagraphic instrument
MCMC	Markov Chain Monte Carlo
Stellar properties	
d	Distance to star
M_*, R_*, L_*	Mass, radius, and luminosity
T_*	Effective temperature
t_*	Age of star and planetary system

Table B.1: (continued)

Symbol	Description
a_{inner}, a_{outer}	Inner and outer edge of the star's habitable zone

Planet properties

Symbol	Description
M_p, R_p, g_p	Mass, radius, and surface gravity
h	Atmospheric scale height
P	Orbital period
a	Semi-major axis
a_{eff}	Semi-major axis scaled by the stellar luminosity, $a_{\mathrm{eff}} = a(L_*/L_\odot)^{-0.5}$
$\cos(i)$	(Cosine of) orbital inclination
b	Transit impact parameter, assuming a circular orbit
δ	Planet transit depth, $\delta = (R_p/R_*)^2$
$\Delta\delta$	Approximate transit depth induced by planet's atmosphere, $\Delta\delta \sim 2(h/R_p)$
ζ	Planet-to-star contrast ratio
R_{est}	Estimated planet radius assuming Earth-like reflectivity (direct imaging only), $R_{\mathrm{est}}/R_\oplus = (\zeta/\zeta_\oplus)^{0.5}(a/1\,\mathrm{AU})$

Simulated survey

Symbol	Description
D_{tel}	Telescope diameter or effective diameter (based on total light-collecting area)
λ_{eff}	Effective wavelength of a spectroscopic measurement
$R_{*,\mathrm{ref}}, T_{*,\mathrm{ref}}$	Radius and effective temperature of the reference star; $(R_{*,\mathrm{ref}}, T_{*,\mathrm{ref}}) = (5777\ \mathrm{K},\ 1\,R_\odot)$ for the imaging survey, $(3000\ \mathrm{K},\ 0.15\,R_\odot)$ for the transit survey
t_i	Amount of time required to characterize the i-th planet in a sample

Table B.1: (continued)

Symbol	Description
t_{ref}	Amount of time required to characterize an Earth twin orbiting the reference star with $a_{\mathrm{eff}} = 1$ AU
t_{total}	Time budget allocated to characterizing planets for a specific spectral feature (may overlap with observations at other wavelengths)
$\zeta_\oplus$	Contrast ratio of the Earth with respect to the Sun, $\zeta_\oplus \approx 10^{-10}$
p_i, w_i	Observing priority and relative weight assigned to each planet, where $p_i = w_i/t_i$

Hypothesis testing

x, y	Independent and dependent variables in the simulated data sets
$h(\vec{\theta}, x)$	Alternative hypothesis describing the relationship between x and y, to be compared to the null hypothesis
$\vec{\theta}$	Set of parameters that define h
$\mathcal{L}(y\|\vec{\theta})$	Likelihood function, described by Equation 6.9 or 6.10
$\Pi(\vec{\theta})$	Prior probability distribution of $\vec{\theta}$, described for each example in Table 6.3
$\mathcal{Z}$	Bayesian evidence in favor of the null or alternative hypothesis, computed by nested sampling

Habitable zone hypothesis

$a_{\mathrm{inner}}, a_{\mathrm{outer}}$	Inner and outer edges of the habitable zone in a_{eff} space (i.e. for a Sun-like star)
$f_{\mathrm{EEC}}^{\mathrm{H_2O}}$	Fraction of EECs with atmospheric water vapor (assumed habitable)

Table B.1: (continued)

Symbol	Description
$f_{\text{non-EEC}}^{\text{H}_2\text{O}}$	Fraction of non-EECs with atmospheric water vapor
Age-oxygen correlation	
f_{life}	Fraction of EECs inhabited by life (regardless of O_2 content)
$t_{1/2}$	Oxygenation timescale; the time required for 50% of inhabited planets to undergo global oxygenation

APPENDIX C

Commissioning the Demo Observatory for Project Nautilus

C.1   Overview

Alongside my thesis research, I have been involved in developing the science case for Project Nautilus, which aims to revolutionize the design of ultra-large space telescopes through the use of multi-order diffractive optical element lenses (MODE lenses; Milster et al., 2020). Project Nautilus is currently developing a 24 cm diameter prototype MODE lens to demonstrate critical manufacturing steps necessary for scaling the technology to larger diameters, such as the precise alignment and fixation of multiple lens segments into a larger whole. Following its assembly, the MODE lens will undergo testing to verify both its predicted optical performance (through optical laboratory tests) and its usefulness for time series astrophysics (through astronomical observations). In this section, I will review my progress in commissioning the demonstration observatory ("demo observatory") where the astronomical testing will be conducted.

C.2   Observatory design

The demo observatory is being commissioned in an 11 foot diameter dome located at the Mount Lemmon SkyCenter north of Tucson, Arizona. Its primary components will consist of a Paramount mount and pier, a custom-machined optical tube assembly (OTA), an Andor camera with a deep-depletion CCD, an off-axis guider telescope, and the MODE prototype lens. A full list of the observatory's major components is given in Table C.1.

Table C.1: Major Components of the Demo Observatory

Component	Manufacturer
11' dome	Ash Manufacturing
Paramount ME robotic mount w/ counterweights	Software Bisque
42" steel pier	. . .
74" OTA	Custom-made
24 cm prototype MODE lens	. . .
Pier-to-floor interface plate	. . .
Telescope-to-mount interface plate	. . .
iKon-M camera	Andor Technology
NexStar 102 SLT (guider telescope)	Celestron
EOS R6 (guider camera)	Canon

C.2.1 OTA and science camera

The OTA consists of a 74"x11" aluminum base, a mounting structure for the MODE lens, and a lightweight cover to mitigate stray light. All of these components were custom-designed and machined for the project. The main science camera is an Andor iKon-M camera with a deep-depletion CCD. This choice of camera was motivated by two considerations: first, the deep-depletion CCD maximizes detector sensitivity, and second, its field of view at the focal plane is wide enough to allow the telescope to observe multiple reference stars simultaneously alongside the primary target. This step is necessary to remove the $\gtrsim 10\%$ level trends that are often present in ground-based time-series observations of stars.

C.2.2 Mount and pier

The mount and pier are capable of supporting the $\sim 50\text{–}60$ kg combined weight of the telescope, lens, and detector, while the mount can accurately track the sky's sidereal motion with an average accumulated error of less than one arcsecond per minute during
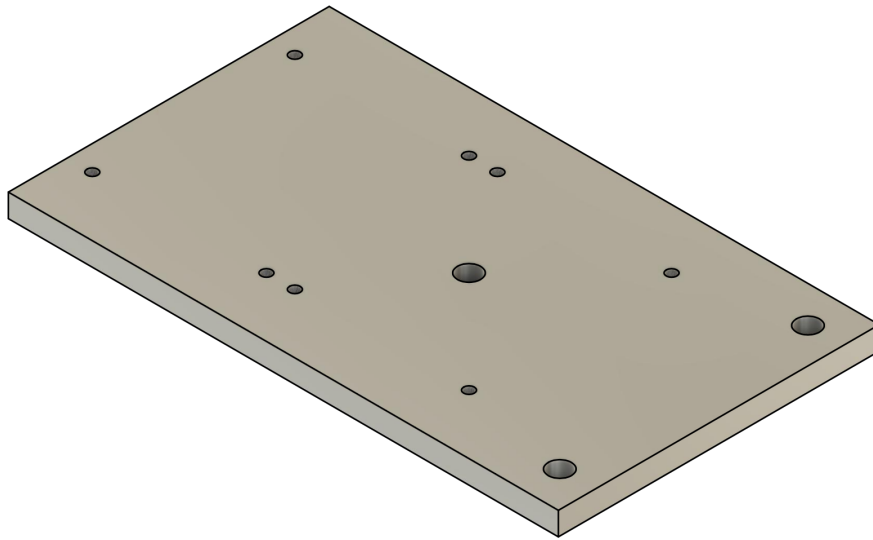
Figure C.1: Drawing of the pier-to-floor interface plate, machined out of steel for rigidity. The three-hole pattern aligns with threaded anchor bolts in the floor of the dome. Two threaded four-hole patterns allow the pier to be offset to the north or south.

unguided observations (as validated in Section C.3.2). The setup also allows for the installation of an off-axis guide telescope, which would further improve the tracking precision. Finally, an extended counterweight shaft with up to 80 lb of weights is available for balancing the setup.

C.2.3 Interface plates

The demo observatory setup combines multiple off-the-shelf and custom-designed components with incompatible interfaces. To resolve this issue, I designed two metal interface plates shown in Figures C.1 and C.2. The first interface plate attaches the pier to the anchor studs in the cement floor of the dome, which were previously installed for a different telescope. The second plate interfaces the Paramount mount with the custom-machined OTA.

Figure C.2: Drawing of the telescope-to-mount interface plate, machined out of aluminum to minimize weight. The six center holes alig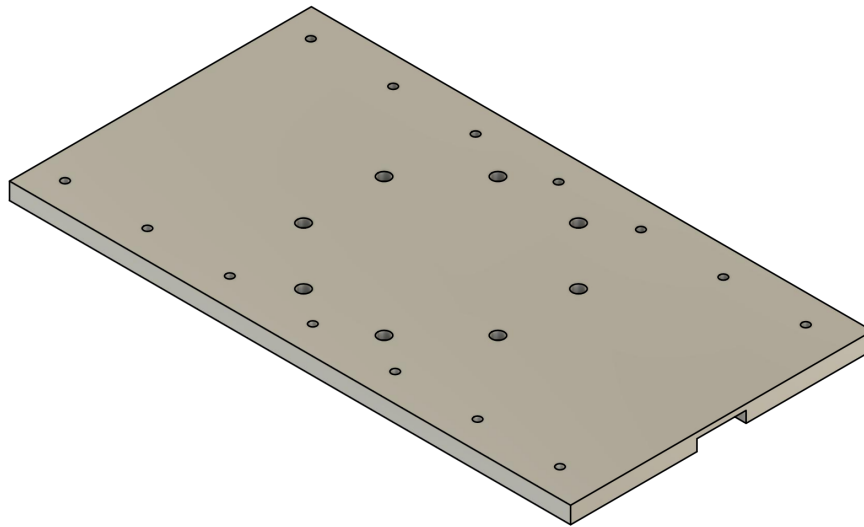n with a similar pattern on top of the mount, while the fourteen holes on the side allow the telescope to be offset by ± 3 inches along the optical axis. The channel on the bottom is used to route power and data cables through the mount to the detector and guide cameras.

C.2.4 Mount, camera, and dome control

The mount and cameras are controlled using TheSkyX[1], a commercial observatory control software package. The mount is connected using a serial-to-USB cable, while the science and guider cameras can be connected using a USB extension cord wired through its base. A small power strip is also wired through the mount to power telescope components.

The laptop can be accessed via remote desktop, allowing for remote control of the mount and camera from elsewhere at the SkyCenter. The dome's two-way rotation motor is operated manually using a lever, so the rotation direction cannot currently be controlled through a digital interface. However, the motor is powered through a network-connected "smart plug" that can be turned on or off remotely via smartphone. By positioning the lever appropriately, a remote observer can activate this plug to rotate the dome in a single direction. This solution is sufficient for the planned MODE prototype tests.

---

[1] https://www.bisque.com/product/theskyx-pro/

Figure C.3: Image of the tripod-mounted setup used for the initial mount tests.

C.3   Testing

The demo observatory is currently undergoing testing to validate the mount's basic performance, measure its sidereal tracking error, and test its guiding capabilities. Successful results from these tests will indicate that the setup is prepared to conduct astronomical tests of the MODE prototype lens once it is completed.

C.3.1   Mount tracking test on tripod

I conducted an initial test of the mount's tracking performance using a small portable tripod. To attach the mount to the tripod, I designed an interface pattern and drilled it into wood. Finally, I secured a 6" refractor to the mount's interface plate to allow for eyepiece observations. The portable setup is pictured in Figure C.3. Testing took place

on two separate nights in a local parking lot, with power provided via portable battery. To begin, I calibrated the mount's pointing solution using 10–20 bright stars spaced across the sky, then applied the polar alignment corrections calculated by the mount's software. The final polar alignment of the mount was accurate to within 20 arcminutes. To test the tracking accuracy, I trained the eyepiece on a target over the eastern horizon and allowed the mount to track it for 90 minutes. Based on the apparent motion of the target from the center of the eyepiece, I estimated the drift rate to be $\sim$ 1–3 arcseconds per minute.

C.3.2   Mount tracking tests on pier

After installing the pier, mount, and interface plates into the dome, I conducted a follow-up test at the SkyCenter to validate and improve upon the previous results. I used the guider telescope and camera for this test, as the primary OTA was not yet installed - this setup is shown in Figure C.4. The use of a camera rather than an eyepiece allowed for precise measurements of stellar positions compared to the previous test, improving both the polar alignment process and tracking error determination.

As before, I calibrated the mount using observations of several bright stars and corrected the polar alignment to within 5 arcminutes accuracy. Next, I trained the telescope on a target over the eastern horizon and allowed it to track for $\sim$ 70 minutes. The results of this test are shown in Figure C.5. The improved polar alignment, combined with new tracking corrections that I activated in the mount software, reduced the drift rate to approximately 0.3 arcseconds per minute.

This tracking error is sufficiently low to conduct stellar time series observations with the MODE prototype lens. Nevertheless, the tracking can be further improved through use of an off-axis guider telescope that measures the cumulative drift of a bright guide star and sends corrective pointing offsets to the mount. Testing this mode, I found the mount was able to stay centered on the star for several minutes with a pointing RMS error of $\sim$ 1.5 arcseconds per 10-second exposure.

Figure C.4: Image of the setup for testing the mount in its permanent dome installation. The 4" guider telescope and camera were used to measure the mount's tracking performance. The blue plate on the floor is the interface plate drawn in Figure C.1.
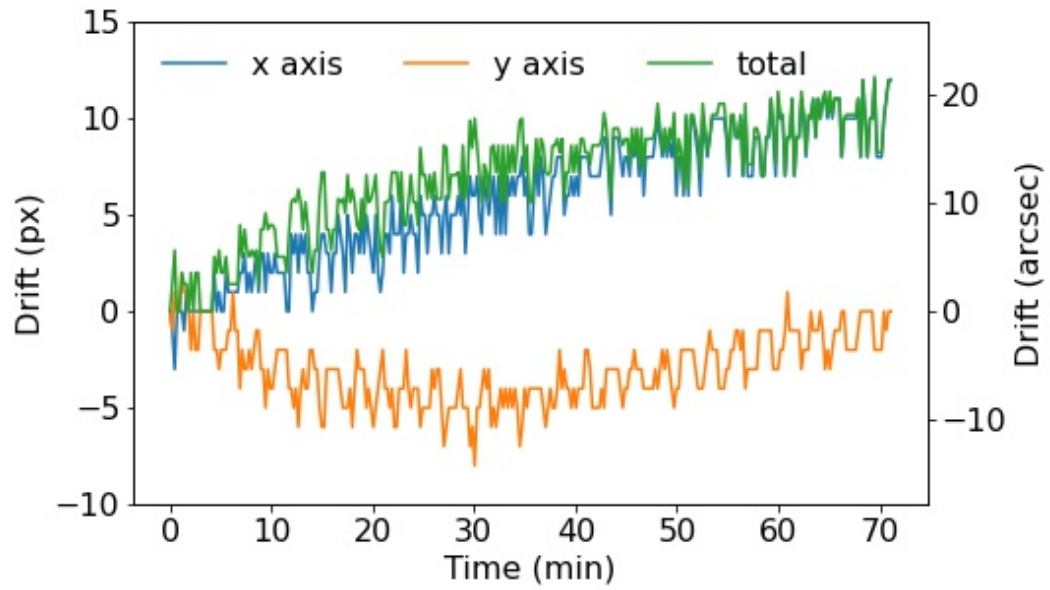
Figure C.5: Results of the tracking test described in Section C.3.2. The accumulated tracking error along each detector axis is plotted as a function of time. The average total drift rate is ~ 0.3 arcseconds per minute.

Figure C.6: The current setup in the demo observatory with the newly-installed OTA. The light baffle and lens mount have been temporarily removed.

C.3.3   Future tests

The OTA has recently been attached to the mount and basic mount capabilities been demonstrated with the heavier load. The current setup is shown in Figure C.6. Next, the mount's performance must be verified again to ensure the tracking is equally capable under increased weight. This will be the final step to validate the mount's performance in advance of the scientific tests. Once the MODE prototype lens has been produced, it will be installed into the telescope along with the science camera and used to conduct multiple observations of eclipsing binary stars and transiting hot Jupiter exoplanets. The quality of the resulting data will be compared both to its ideal performance (i.e. as determined by its light-collecting area) as well as data from a nearby robotic telescope that is frequently used for similar observations. Finally, the peer-reviewed publication of these test results (alongside laboratory results) will close out this stage of MODE lens development.

C.4   Acknowledgements

REFERENCES

Aizawa, M., H. Kawahara, and S. Fan (2020). Global Mapping of an Exo-Earth Using Sparse Modeling. *ApJ*, **896**(1), 22.

Akeson, R., L. Armus, E. Bachelet, et al. (2019). The Wide Field Infrared Survey Telescope: 100 Hubbles for the 2020s. *arXiv e-prints*, arXiv:1902.05569.

ALMA Partnership, C. L. Brogan, L. M. Pérez, et al. (2015). The 2014 ALMA Long Baseline Campaign: First Results from High Angular Resolution Observations toward the HL Tau Region. *ApJ*, **808**(1), L3.

Anbar, A. D., Y. Duan, T. W. Lyons, et al. (2007). A whiff of oxygen before the great oxidation event? *Science*, **317**(5846), pp. 1903–1906.

Angel, J. R., A. Y. Cheng, and N. J. Woolf (1986). A space telescope for infrared spectroscopy of Earth-like planets. *Nature*, **322**(6077), pp. 341–343.

Anglada-Escudé, G., P. J. Amado, J. Barnes, et al. (2016). A terrestrial planet candidate in a temperate orbit around Proxima Centauri. *Nature*, **536**, pp. 437–440.

Apai, D., A. Bixel, B. V. Rackham, et al. (2019). Nautilus: A Very Large-Aperture, Ultralight Space Telescope for Exoplanet Exploration, Time-domain Astrophysics, and Faint Objects. *BAAS*, **51**, 141.

Apai, D., F. Ciesla, G. D. Mulders, et al. (2018a). A comprehensive understanding of planet formation is required for assessing planetary habitability and for the search for life. *arXiv e-prints*, arXiv:1803.08682.

Apai, D., N. Cowan, R. Kopparapu, et al. (2017). Exploring Other Worlds: Science Questions for Future Direct Imaging Missions (EXOPAG SAG15 Report). *arXiv e-prints*, arXiv:1708.02821.

Apai, D., T. D. Milster, D. W. Kim, et al. (2019). A Thousand Earths: A Very Large Aperture, Ultralight Space Telescope Array for Atmospheric Biosignature Surveys. *AJ*, **158**(2), p. 83.

Apai, D., T. D. Milster, D. W. Kim, et al. (2019). Nautilus Observatory: a space telescope array based on very large aperture ultralight diffractive optical elements. In *Astronomical Optics: Design, Manufacture, and Test of Space and Ground Systems II*, volume 11116 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 1111608.

Apai, D., B. V. Rackham, M. S. Giampapa, et al. (2018b). Understanding Stellar Contamination in Exoplanet Transmission Spectra as an Essential Step in Small Planet Characterization. *arXiv e-prints*, arXiv:1803.08708.

Apai, D., J. Radigan, E. Buenzli, et al. (2013). HST Spectral Mapping of L/T Transition Brown Dwarfs Reveals Cloud Thickness Variations. *ApJ*, **768**, 121.

Arney, G., S. D. Domagal-Goldman, and V. S. Meadows (2018). Organic Haze as a Biosignature in Anoxic Earth-like Atmospheres. *Astrobiology*, **18**(3), pp. 311–329.

Arney, G., S. D. Domagal-Goldman, V. S. Meadows, et al. (2016). The Pale Orange Dot: The Spectrum and Habitability of Hazy Archean Earth. *Astrobiology*, **16**(11), pp. 873–899.

Arney, G. N., V. S. Meadows, S. D. Domagal-Goldman, et al. (2017). Pale Orange Dots: The Impact of Organic Haze on the Habitability and Detectability of Earthlike Exoplanets. *ApJ*, **836**(1), p. 49.

Artigau, E., R. A. Bernstein, T. Brandt, et al. (2018). Direct Imaging in Reflected Light: Characterization of Older, Temperate Exoplanets With 30-m Telescopes. *arXiv e-prints*, arXiv:1808.09632.

Ballard, S. (2019). Predicted Number, Multiplicity, and Orbital Dynamics of TESS M-dwarf Exoplanets. *AJ*, **157**(3), 113.

Baran, A. S., L. Fox-Machado, J. Lykke, et al. (2011a). Mt. Suhora Survey - Searching for Pulsating M Dwarfs in Kepler Public Dataset. *AcA*, **61**(4), pp. 325–343.

Baran, A. S., M. Winiarski, J. Krzesiński, et al. (2011b). Mt. Suhora Survey - Searching for Pulsating M Dwarfs. I. *AcA*, **61**(1), pp. 37–58.

Baran, A. S., M. Winiarski, M. Siwak, et al. (2013). Mt. Suhora Survey - Searching for Pulsating M Dwarfs. III. *AcA*, **63**(1), pp. 41–52.

Barclay, T., J. Pepper, and E. V. Quintana (2018). A Revised Exoplanet Yield from the Transiting Exoplanet Survey Satellite (TESS). *ApJS*, **239**(1), 2.

Barnes, R., R. Deitrick, R. Luger, et al. (2016). The Habitability of Proxima Centauri b I: Evolutionary Scenarios. *arXiv e-prints*, arXiv:1608.06919.

Barry, R., J. Kruk, J. Anderson, et al. (2011). The exoplanet microlensing survey by the proposed WFIRST Observatory. In *Techniques and Instrumentation for Detection of Exoplanets V*, volume 8151 of *Proc. SPIE*, p. 81510L.

Batalha, N. E., A. J. R. W. Smith, N. K. Lewis, et al. (2018). Color Classification of Extrasolar Giant Planets: Prospects and Cautions. *AJ*, **156**(4), 158.

Bean, J. L., D. S. Abbot, and E. M.-R. Kempton (2017). A Statistical Comparative Planetology Approach to the Hunt for Habitable Exoplanets and Life Beyond the Solar System. *ApJ*, **841**(2), p. L24.

Beerer, I. M., H. A. Knutson, A. Burrows, et al. (2011). Secondary Eclipse Photometry of WASP-4b with Warm Spitzer. *ApJ*, **727**, 23.

Béky, B., D. M. Kipping, and M. J. Holman (2014). SPOTROD: a semi-analytic model for transits of spotted stars. *MNRAS*, **442**, pp. 3686–3699.

Belikov, R., E. Bendek, S. Thomas, et al. (2015). How to directly image a habitable planet around Alpha Centauri with a ~30-45cm space telescope. In *Techniques and Instrumentation for Detection of Exoplanets VII*, volume 9605 of *Proc. SPIE*, p. 960517.

Benneke, B. and S. Seager (2013). How to Distinguish between Cloudy Mini-Neptunes and Water/Volatile-dominated Super-Earths. *ApJ*, **778**, 153.

Benneke, B., I. Wong, C. Piaulet, et al. (2019). Water Vapor and Clouds on the Habitable-zone Sub-Neptune Exoplanet K2-18b. *ApJL*, **887**(1), L14.

Berdiñas, Z. M., C. Rodríguez-López, P. J. Amado, et al. (2017). High-cadence spectroscopy of M-dwarfs - II. Searching for stellar pulsations with HARPS. *MNRAS*, **469**(4), pp. 4268–4282.

Bétrémieux, Y. and M. R. Swain (2017). An analytical formalism accounting for clouds and other 'surfaces' for exoplanet transmission spectroscopy. *MNRAS*, **467**, pp. 2834–2844.

Biller, B. A., J. Vos, M. Bonavita, et al. (2015). Variability in a Young, L/T Transition Planetary-mass Object. *ApJL*, **813**, L23.

Bixel, A. and D. Apai (2017). Probabilistic Constraints on the Mass and Composition of Proxima b. *ApJL*, **836**(2), L31.

Bixel, A. and D. Apai (2020a). Identifying Exo-Earth Candidates in Direct Imaging Data through Bayesian Classification. *AJ*, **159**(1), 3.

Bixel, A. and D. Apai (2020b). Testing Earthlike Atmospheric Evolution on Exo-Earths through Oxygen Absorption: Required Sample Sizes and the Advantage of Age-based Target Selection. *ApJ*, **896**(2), 131.

Bixel, A. and D. Apai (2021). Bioverse: A Simulation Framework to Assess the Statistical Power of Future Biosignature Surveys. *The Astronomical Journal*, **161**(5), p. 228.

Bixel, A. and D. Apai (2021). Bioverse: a simulation framework to assess the statistical power of future biosignature surveys, 1.0, Zenodo, doi:10.5281/zenodo.4460666.

Bixel, A., B. V. Rackham, D. Apai, et al. (2019). ACCESS: Ground-based Optical Transmission Spectroscopy of the Hot Jupiter WASP-4b. *AJ*, **157**(2), 68.

Bolmont, E., A.-S. Libert, J. Leconte, and F. Selsis (2016). Habitability of planets on eccentric orbits: Limits of the mean flux approximation. *A&A*, **591**, A106.

Borucki, W. J., D. Koch, G. Basri, et al. (2010). Kepler Planet-Detection Mission: Introduction and First Results. *Science*, **327**(5968), p. 977.

Bouma, L. G., J. N. Winn, J. Kosiarek, and P. R. McCullough (2017). Planet Detection Simulations for Several Possible TESS Extended Missions. *arXiv e-prints*, arXiv:1705.08891.

Bowler, B. P. (2016). Imaging Extrasolar Giant Planets. *PASP*, **128**(968), p. 102001.

Brocks, J. J., G. A. Logan, R. Buick, and R. E. Summons (1999). Archean molecular fossils and the early rise of eukaryotes. *Science*, **285**(5430), pp. 1033–1036.

Broeg, C., A. Fortier, D. Ehrenreich, et al. (2013). CHEOPS: A transit photometry mission for ESA's small mission programme. In *European Physical Journal Web of Conferences*, volume 47 of *European Physical Journal Web of Conferences*, p. 03005.

Brown, T. M. (2001). Transmission Spectra as Diagnostics of Extrasolar Giant Planet Atmospheres. *ApJ*, **553**, pp. 1006–1026.

Brugger, B., O. Mousis, M. Deleuil, and J. I. Lunine (2016). Possible Internal Structures and Compositions of Proxima Centauri b. *ApJL*, **831**, L16.

Bryson, S., M. Kunimoto, R. K. Kopparapu, et al. (2021). The Occurrence of Rocky Habitable-zone Planets around Solar-like Stars from Kepler Data. *AJ*, **161**(1), 36.

Buccino, A. P., G. A. Lemarchand, and P. J. Mauas (2007). UV habitable zones around M stars. *Icarus*, **192**(2), pp. 582–587.

Buchner, J., A. Georgakakis, K. Nandra, et al. (2014). X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue. *A&A*, **564**, A125.

Buenzli, E., D. Apai, J. Radigan, et al. (2014). Brown Dwarf Photospheres are Patchy: A Hubble Space Telescope Near-infrared Spectroscopic Survey Finds Frequent Low-level Variability. *ApJ*, **782**, 77.

Burgasser, A. J. and E. E. Mamajek (2017). On the Age of the TRAPPIST-1 System. *ApJ*, **845**(2), 110.

Caldeira, K. and J. F. Kasting (1992). The life span of the biosphere revisited. *Nature*, **360**(6406), pp. 721–723.

Carrión-González, Ó., A. García Muñoz, J. Cabrera, et al. (2020). Directly imaged exoplanets in reflected starlight: the importance of knowing the planet radius. *A&A*, **640**, A136.

Carter, J. A. and J. N. Winn (2009). Parameter Estimation from Time-series Data with Correlated Errors: A Wavelet-based Method and its Application to Transit Light Curves. *ApJ*, **704**, pp. 51–67.

Cash, W. (2006). Detection of Earth-like planets around nearby stars using a petal-shaped occulter. *Nature*, **442**(7098), pp. 51–53.

Catling, D. C., C. R. Glein, K. J. Zahnle, and C. P. McKay (2005). Why O_2 Is Required by Complex Life on Habitable Planets and the Concept of Planetary "Oxygenation Time". *Astrobiology*, **5**(3), pp. 415–438.

Catling, D. C., J. Krissansen-Totton, N. Y. Kiang, et al. (2018). Exoplanet Biosignatures: A Framework for Their Assessment. *Astrobiology*, **18**(6), pp. 709–738.

Catling, D. C. and K. J. Zahnle (2020). The Archean atmosphere. *Science Advances*, **6**(9), p. eaax1420.

Catling, D. C., K. J. Zahnle, and C. P. McKay (2001). Biogenic methane, hydrogen escape, and the irreversible oxidation of early earth. *Science*, **293**(5531), pp. 839–843.

Cauley, P. W., C. Kuckein, S. Redfield, et al. (2018). The Effects of Stellar Activity on Optical High-resolution Exoplanet Transmission Spectra. *AJ*, **156**, 189.

Chabrier, G. (2003). Galactic Stellar and Substellar Initial Mass Function. *PASP*, **115**(809), pp. 763–795.

Chaplin, W. J., S. Basu, D. Huber, et al. (2014). Asteroseismic Fundamental Properties of Solar-type Stars Observed by the NASA Kepler Mission. *ApJS*, **210**(1), 1.

Charbonneau, D., T. M. Brown, R. W. Noyes, and R. L. Gilliland (2002). Detection of an Extrasolar Planet Atmosphere. *ApJ*, **568**, pp. 377–384.

Checlair, J., D. S. Abbot, R. J. Webber, et al. (2019). A Statistical Comparative Planetology Approach to Maximize the Scientific Return of Future Exoplanet Characterization Efforts. *BAAS*, **51**(3), 404.

Checlair, J. H., G. L. Villanueva, B. P. C. Hayworth, et al. (2021). Probing the Capability of Future Direct-imaging Missions to Spectrally Constrain the Frequency of Earth-like Planets. *AJ*, **161**(3), 150.

Chen, H. and L. A. Rogers (2016). Evolutionary Analysis of Gaseous Sub-Neptune-mass Planets with MESA. *ApJ*, **831**, 180.

Chen, J. and D. Kipping (2017). Probabilistic Forecasting of the Masses and Radii of Other Worlds. *ApJ*, **834**, 17.

Claire, M. W., D. C. Catling, and K. J. Zahnle (2006). Biogeochemical modelling of the rise in atmospheric oxygen. *Geobiology*, **4**(4), pp. 239–269.

Close, L. M. (2016). A review of astronomical science with visible light adaptive optics. In Marchetti, E., L. M. Close, and J.-P. Véran (eds.) *Adaptive Optics Systems V*, volume 9909 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 99091E.

Coleman, G. A. L., R. P. Nelson, S. J. Paardekooper, et al. (2017). Exploring plausible formation scenarios for the planet candidate orbiting Proxima Centauri. *MNRAS*.

Cowan, N. B., E. Agol, V. S. Meadows, et al. (2009). Alien Maps of an Ocean-Bearing World. *ApJ*, **700**(2), p. 915.

Cowan, N. B. and Y. Fujii (2018). Mapping Exoplanets. In *Handbook of Exoplanets*, pp. 1469–1484. Springer International Publishing, Cham.

Creevey, O. L., T. S. Metcalfe, M. Schultheis, et al. (2017). Characterizing solar-type stars from full-length Kepler data sets using the Asteroseismic Modeling Portal. *A&A*, **601**, A67.

Crill, B. P. and N. Siegler (2017). Space technology for directly imaging and characterizing exo-Earths. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 10398 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 103980H.

de Wit, J., H. R. Wakeford, N. K. Lewis, et al. (2018). Atmospheric reconnaissance of the habitable-zone Earth-sized planets orbiting TRAPPIST-1. *Nature Astronomy*, **2**, pp. 214–219.

Deming, D., A. Wilkins, P. McCullough, et al. (2013). Infrared Transmission Spectroscopy of the Exoplanets HD 209458b and XO-1b Using the Wide Field Camera-3 on the Hubble Space Telescope. *ApJ*, **774**, 95.

Des Marais, D. J., M. O. Harwit, K. W. Jucks, et al. (2002). Remote sensing of planetary properties and biosignatures on extrasolar terrestrial planets. *Astrobiology*, **2**(2), pp. 153–181.

Diamond-Lowe, H., K. B. Stevenson, J. L. Bean, et al. (2014). New Analysis Indicates No Thermal Inversion in the Atmosphere of HD 209458b. *ApJ*, **796**, 66.

Domagal-Goldman, S. D., V. S. Meadows, M. W. Claire, and J. F. Kasting (2011). Using biogenic sulfur gases as remotely detectable biosignatures on anoxic planets. *Astrobiology*, **11**(5), pp. 419–441.

Domagal-Goldman, S. D., A. Segura, M. W. Claire, et al. (2014). Abiotic ozone and oxygen in atmospheres similar to prebiotic earth. *ApJ*, **792**(2).

Dorn, C., L. Noack, and A. B. Rozel (2018). Outgassing on stagnant-lid super-Earths. *A&A*, **614**, A18.

Dressing, C., C. C. Stark, P. Plavchan, and E. Lopez (2019). Ground-Based Radial Velocity as Critical Support for Future NASA Earth-Finding Missions. *BAAS*, **51**(3), 268.

Dressler, A., B. Bigelow, T. Hare, et al. (2011). IMACS: The Inamori-Magellan Areal Camera and Spectrograph on Magellan-Baade. *PASP*, **123**, p. 288.

Driscoll, P. E. and R. Barnes (2015). Tidal heating of earth-like exoplanets around M Stars: Thermal, magnetic, and orbital evolutions. *Astrobiology*, **15**(9), pp. 739–760.

Ertel, S., D. Defrère, P. Hinz, et al. (2018). The HOSTS Survey—Exozodiacal Dust Measurements for 30 Stars. *AJ*, **155**, 194.

ESA (2017). PLATO Definition Study Report, https://sci.esa.int/web/plato.

Espinoza, N. and A. Jordán (2015). Limb darkening and exoplanets: testing stellar model atmospheres and identifying biases in transit parameters. *MNRAS*, **450**, pp. 1879–1899.

Espinoza, N., B. V. Rackham, A. Jordán, et al. (2019). ACCESS: a featureless optical transmission spectrum for WASP-19b from Magellan/IMACS. *MNRAS*, **482**, pp. 2065–2087.

Fan, S., C. Li, J.-Z. Li, et al. (2019). Earth as an Exoplanet: A Two-dimensional Alien Map. *ApJ*, **882**(1), p. L1.

Fantin, N. J., P. Côté, A. W. McConnachie, et al. (2019). The Canada-France Imaging Survey: Reconstructing the Milky Way Star Formation History from Its White Dwarf Population. *ApJ*, **887**(2), 148.

Farr, B., W. M. Farr, N. B. Cowan, et al. (2018). exocartographer: A Bayesian Framework for Mapping Exoplanets in Reflected Light . *AJ*, **156**(4), p. 146.

Fauchez, T. J., M. Turbet, G. L. Villanueva, et al. (2019). Impact of Clouds and Hazes on the Simulated JWST Transmission Spectra of Habitable Zone Planets in the TRAPPIST-1 System. *ApJ*, **887**(2), 194.

Fawdon, P., S. Gupta, J. M. Davis, et al. (2018). The Hypanis Valles delta: The last highstand of a sea on early Mars? *Earth and Planetary Science Letters*, **500**, pp. 225–241.

Ford, E. B., S. Seager, and E. L. Turner (2001). Characterization of extrasolar terrestrial planets from diurnal photometric variability. *Nature*, **412**(6850), pp. 885–887.

Foreman-Mackey, D., D. W. Hogg, D. Lang, and J. Goodman (2013). emcee: The MCMC Hammer. *PASP*, **125**(925), p. 306.

Fujii, Y., D. Angerhausen, R. Deitrick, et al. (2018). Exoplanet Biosignatures: Observational Prospects. *Astrobiology*, **18**, pp. 739–778.

Fulton, B. J. and E. A. Petigura (2018). The California-Kepler Survey. VII. Precise Planet Radii Leveraging Gaia DR2 Reveal the Stellar Mass Dependence of the Planet Radius Gap. *AJ*, **156**(6), 264.

Fulton, B. J., E. A. Petigura, A. W. Howard, et al. (2017). The California-Kepler Survey. III. A Gap in the Radius Distribution of Small Planets. *AJ*, **154**, 109.

Gaia Collaboration (2018). VizieR Online Data Catalog: Gaia DR2 (Gaia Collaboration, 2018). *VizieR Online Data Catalog*, I/345.

Gale, J. and A. Wandel (2016). The potential of planets orbiting red dwarf stars to support oxygenic photosynthesis and complex life. *International Journal of Astrobiology*, **16**(1), pp. 1–9.

Gao, P., R. Hu, T. D. Robinson, et al. (2015). Stability of Co2 Atmospheres on Desiccated M Dwarf Exoplanets. *ApJ*, **806**(2).

Gaudi, B. S., S. Seager, B. Mennesson, et al. (2020). The Habitable Exoplanet Observatory (HabEx) Mission Concept Study Final Report. *arXiv e-prints*, arXiv:2001.06683.

Gillon, M., B. Smalley, L. Hebb, et al. (2009). Improved parameters for the transiting hot Jupiters WASP-4b and WASP-5b. *A&A*, **496**, pp. 259–267.

Gillon, M., A. H. M. J. Triaud, B.-O. Demory, et al. (2017). Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1. *Nature*, **542**(7642), pp. 456–460.

Gilmozzi, R. and J. Spyromilio (2007). The European Extremely Large Telescope (E-ELT). *The Messenger*, **127**, p. 11.

Goldblatt, C. (2016). Tutorial models of the climate and habitability of Proxima Centauri b: a thin atmosphere is sufficient to distribute heat given low stellar flux. *ArXiv e-prints*.

Guimond, C. M. and N. B. Cowan (2018). The Direct Imaging Search for Earth 2.0: Quantifying Biases and Planetary False Positives. *AJ*, **155**, 230.

Guimond, C. M. and N. B. Cowan (2019). Three Direct Imaging Epochs Could Constrain the Orbit of Earth 2.0 inside the Habitable Zone. *AJ*, **157**(5), 188.

Gupta, A. and H. E. Schlichting (2019). Sculpting the valley in the radius distribution of small exoplanets as a by-product of planet formation: the core-powered mass-loss mechanism. *MNRAS*, **487**, pp. 24–33.

Hansen, C. J., S. D. Kawaler, and V. Trimble (2004). *Stellar interiors : physical principles, structure, and evolution*.

Haqq-Misra, J. (2019). Does the Evolution of Complex Life Depend on the Stellar Spectral Energy Distribution? *Astrobiology*, **19**(10), pp. 1292–1299.

Haqq-Misra, J. and R. K. Kopparapu (2018). The Drake Equation as a Function of Spectral Type and Time. In *Habitability of the Universe Before Earth*, pp. 307–319. Elsevier.

Hardegree-Ullman, K. K., M. C. Cushing, P. S. Muirhead, and J. L. Christiansen (2019). Kepler Planet Occurrence Rates for Mid-type M Dwarfs as a Function of Spectral Type. *AJ*, **158**(2), 75.

Harman, C. E., R. Felton, R. Hu, et al. (2018). Abiotic O 2 Levels on Planets around F, G, K, and M Stars: Effects of Lightning-produced Catalysts in Eliminating Oxygen False Positives . *ApJ*, **866**(1), p. 56.

Harman, C. E., E. W. Schwieterman, J. C. Schottelkotte, and J. F. Kasting (2015). Abiotic O2 levels on planets around F, G, K, and M stars: Possible false positives for life? *ApJ*, **812**(2).

Hart, M. H. (1979). Habitable Zones about Main Sequence Stars. *Icarus*, **37**(1), pp. 351–357.

Heng, K. and D. Kitzmann (2017). The theory of transmission spectra revisited: a semi-analytical method for interpreting WFC3 data and an unresolved challenge. *MNRAS*, **470**, pp. 2972–2981.

Ho, S. and E. L. Turner (2011). The Posterior Distribution of sin(i) Values for Exoplanets with M_T sin(i) Determined from Radial Velocity Data. *ApJ*, **739**, 26.

Howard, W. S., M. A. Tilley, H. Corbett, et al. (2018). The First Naked-eye Super fl are Detected from Proxima Centauri. *ApJL*, **860**(2), p. L30.

Hoyer, S., M. López-Morales, P. Rojo, et al. (2013). TraMoS project - III. Improved physical parameters, timing analysis and starspot modelling of the WASP-4b exoplanet system from 38 transit observations. *MNRAS*, **434**, pp. 46–58.

Hsu, D. C., E. B. Ford, and R. Terrien (2020). Occurrence Rates of Planets Orbiting M Stars: Applying ABC to Kepler DR25, Gaia DR2, and 2MASS Data. *MNRAS*.

Hu, R., L. Peterson, and E. T. Wolf (2020). O_2- and CO-rich Atmospheres for Potentially Habitable Environments on TRAPPIST-1 Planets. *ApJ*, **888**(2), 122.

Hu, R., S. Seager, and W. Bains (2012). Photochemistry in terrestrial exoplanet atmospheres. I. Photochemistry model and benchmark cases. *ApJ*, **761**(2).

Hu, R., S. Seager, and Y. L. Yung (2015). Helium Atmospheres on Warm Neptune- and Sub-Neptune-sized Exoplanets and Applications to GJ 436b. *ApJ*, **807**(1), 8.

Hubbard, W. B., J. J. Fortney, J. I. Lunine, et al. (2001). Theory of Extrasolar Giant Planet Transits. *ApJ*, **560**, pp. 413–419.

Huitson, C. M., J. M. Désert, J. L. Bean, et al. (2017). Gemini/GMOS Transmission Spectral Survey: Complete Optical Transmission Spectrum of the Hot Jupiter WASP-4b. *AJ*, **154**, 95.

Hunten, D. M. and T. M. Donahue (1976). Hydrogen Loss from the Terrestrial Planets. *Annual Review of Earth and Planetary Sciences*, **4**(1), pp. 265–292.

Husser, T.-O., S. Wende-von Berg, S. Dreizler, et al. (2013). A new extensive library of PHOENIX stellar atmospheres and synthetic spectra. *A&A*, **553**, A6.

Iyer, A. R. and M. R. Line (2020). The Influence of Stellar Contamination on the Interpretation of Near-infrared Transmission Spectra of Sub-Neptune Worlds around M-dwarfs. *ApJ*, **889**(2), 78.

Jakosky, B. M., J. M. Grebowsky, J. G. Luhmann, et al. (2015). MAVEN observations of the response of Mars to an interplanetary coronal mass ejection. *Science*, **350**(6261), pp. 1–8.

Jakosky, B. M., M. Slipski, M. Benna, et al. (2017). Mars' atmospheric history derived from upper-atmosphere measurements of 38Ar/36Ar. *Science*, **355**(6332), pp. 1408–1410.

James, T. and R. Hu (2018). Photochemical Oxygen in Non-1-bar CO 2 Atmospheres of Terrestrial Exoplanets . *ApJ*, **867**(1), p. 17.

Johns, M., P. McCarthy, K. Raybould, et al. (2012). Giant Magellan Telescope: overview. In Stepp, L. M., R. Gilmozzi, and H. J. Hall (eds.) *Ground-based and Airborne Telescopes IV*, volume 8444 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 84441H.

Jordán, A., N. Espinoza, M. Rabus, et al. (2013). A Ground-based Optical Transmission Spectrum of WASP-6b. *ApJ*, **778**, 184.

Juanola-Parramon, R., N. T. Zimmerman, L. Pueyo, et al. (2019). The LUVOIR Extreme Coronagraph for Living Planetary Systems (ECLIPS) II. Performance evaluation, aberration sensitivity analysis and exoplanet detection simulations. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 11117 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 1111702.

Kammerer, J. and S. P. Quanz (2018). Simulating the exoplanet yield of a space-based mid-infrared interferometer based on Kepler statistics. *A&A*, **609**, A4.

Kane, S. R., D. R. Ciardi, D. M. Gelino, and K. von Braun (2012). The exoplanet eccentricity distribution from Kepler planet candidates. *MNRAS*, **425**, pp. 757–762.

Kasper, M., J.-L. Beuzit, C. Verinaud, et al. (2010). EPICS: direct imaging of exoplanets with the E-ELT. In *Ground-based and Airborne Instrumentation for Astronomy III*, volume 7735 of *Proc. SPIE*, pp. 77352E–77352E–9.

Kasting, J. F., D. P. Whitmire, and R. T. Reynolds (1993). Habitable Zones around Main Sequence Stars. *Icarus*, **101**(1), pp. 108–128.

Kataria, T., D. K. Sing, N. K. Lewis, et al. (2016). The Atmospheric Circulation of a Nine-hot-Jupiter Sample: Probing Circulation and Chemistry over a Wide Phase Space. *ApJ*, **821**, 9.

Kawashima, Y. and S. Rugheimer (2019). Theoretical Reflectance Spectra of Earth-like Planets through Their Evolutions: Impact of Clouds on the Detectability of Oxygen, Water, and Methane with Future Direct Imaging Missions. *AJ*, **157**(5), p. 213.

Kayhan, C., M. Yıldız, and Z. Çelik Orhan (2019). Asteroseismic investigation of 20 planet and planet-candidate host stars. *MNRAS*, **490**(2), pp. 1509–1517.

Kendall, B., C. T. Reinhard, T. W. Lyons, et al. (2010). Pervasive oxygenation along late Archaean ocean margins. *Nature Geoscience*, **3**(9), pp. 647–652.

Kervella, P., F. Thévenin, and C. Lovis (2017). Proxima's orbit around α Centauri. *A&A*, **598**, L7.

Kiang, N. Y., A. Segura, G. Tinetti, et al. (2007). Spectral signatures of photosynthesis. II. Coevolution with other stars and the atmosphere on extrasolar worlds. *Astrobiology*, **7**(1), pp. 252–274.

Kipping, D. M. (2013a). Efficient, uninformative sampling of limb darkening coefficients for two-parameter laws. *MNRAS*, **435**, pp. 2152–2160.

Kipping, D. M. (2013b). Parametrizing the exoplanet eccentricity distribution with the beta distribution. *MNRAS*, **434**, pp. L51–L55.

Kochanek, C. S., B. J. Shappee, K. Z. Stanek, et al. (2017). The All-Sky Automated Survey for Supernovae (ASAS-SN) Light Curve Server v1.0. *PASP*, **129**(10), p. 104502.

Kodama, T., A. Nitta, H. Genda, et al. (2018). Dependence of the Onset of the Runaway Greenhouse Effect on the Latitudinal Surface Water Distribution of Earth-Like Planets. *Journal of Geophysical Research (Planets)*, **123**(2), pp. 559–574.

Komacek, T. D. and D. S. Abbot (2019). The Atmospheric Circulation and Climate of Terrestrial Planets Orbiting Sun-like and M Dwarf Stars over a Broad Range of Planetary Parameters. *ApJ*, **871**(2), 245.

Komacek, T. D., T. J. Fauchez, E. T. Wolf, and D. S. Abbot (2020). Clouds will Likely Prevent the Detection of Water Vapor in JWST Transmission Spectra of Terrestrial Exoplanets. *ApJL*, **888**(2), L20.

Kopparapu, R. K., E. Hébrard, R. Belikov, et al. (2018). Exoplanet Classification and Yield Estimates for Direct Imaging Missions. *ApJ*, **856**(2), p. 122.

Kopparapu, R. K., R. Ramirez, J. F. Kasting, et al. (2013). Habitable zones around main-sequence stars: New estimates. *ApJ*, **765**(2), p. 131.

Kopparapu, R. K., R. M. Ramirez, J. Schottelkotte, et al. (2014). Habitable zones around main-sequence stars: Dependence on planetary mass. *ApJL*, **787**(2), p. L29.

Kreidberg, L., J. L. Bean, J. M. Désert, et al. (2014). Clouds in the atmosphere of the super-Earth exoplanet GJ 1214b. *Nature*, **505**(7481), pp. 69–72.

Krissansen-Totton, J., S. Olson, and D. C. Catling (2018). Disequilibrium biosignatures over Earth history and implications for detecting exoplanet life. *Science Advances*, **4**(1).

Krissansen-Totton, J., E. W. Schwieterman, B. Charnay, et al. (2016). Is the Pale Blue Dot Unique? Optimized Photometric Bands for Identifying Earth-like Exoplanets. *ApJ*, **817**(1), 31.

Krzesinski, J., A. S. Baran, M. Winiarski, et al. (2012). Mt. Suhora Survey - Searching for Pulsating M Dwarfs. II. *AcA*, **62**(2), pp. 201–212.

Kurzweil, F., M. Claire, C. Thomazo, et al. (2013). Atmospheric sulfur rearrangement 2.7 billion years ago: Evidence for oxygenic photosynthesis. *Earth and Planetary Science Letters*, **366**, pp. 17–26.

Lehmer, O. R., D. C. Catling, and J. Krissansen-Totton (2020). Carbonate-silicate cycle predictions of Earth-like planetary climates and testing the habitable zone concept. *Nature Communications*, **11**(1).

Lehmer, O. R., D. C. Catling, M. N. Parenteau, and T. M. Hoehler (2018). The Productivity of Oxygenic Photosynthesis around Cool, M Dwarf Stars. *ApJ*, **859**(2), p. 171.

Ligi, R., O. Creevey, D. Mourard, et al. (2016). Radii, masses, and ages of 18 bright stars using interferometry and new estimations of exoplanetary parameters. *A&A*, **586**, A94.

Lingam, M. and A. Loeb (2018). Physical constraints for the evolution of life on exoplanets. *Reviews of Modern Physics*, **91**(2), p. 021002.

Lingam, M. and A. Loeb (2019). Photosynthesis on habitable planets around low-mass stars. *MNRAS*, **485**(4), pp. 5924–5928.

Lopez, E. D. and J. J. Fortney (2014). Understanding the Mass-Radius Relation for Subneptunes: Radius as a Proxy for Composition. *ApJ*, **792**(1), 1.

Lovelock, J. E. (1965). A physical basis for life detection experiments. *Nature*, **207**(4997), pp. 568–570.

Lovis, C., I. Snellen, D. Mouillet, et al. (2017). Atmospheric characterization of Proxima b by coupling the SPHERE high-contrast imager to the ESPRESSO spectrograph. *A&A*, **599**, A16.

Luger, R. and R. Barnes (2015). Extreme water loss and abiotic O 2 buildup on planets throughout the habitable zones of m dwarfs. *Astrobiology*, **15**(2), pp. 119–143.

Luger, R., J. Lustig-Yaeger, D. P. Fleming, et al. (2017). The Pale Green Dot: A Method to Characterize Proxima Centauri b Using Exo-Aurorae. *ApJ*, **837**(1), 63.

Lund, M. N., E. Knudstrup, V. Silva Aguirre, et al. (2019). Asteroseismology of the Multiplanet System K2-93. *AJ*, **158**(6), 248.

Lunine, J. I. (2013). *Earth : evolution of a habitable world*. Cambridge University Press.

Lustig-Yaeger, J., V. S. Meadows, and A. P. Lincowski (2019a). A Mirage of the Cosmic Shoreline: Venus-like Clouds as a Statistical False Positive for Exoplanet Atmospheric Erosion. *ApJL*, **887**(1), L11.

Lustig-Yaeger, J., V. S. Meadows, and A. P. Lincowski (2019b). The Detectability and Characterization of the TRAPPIST-1 Exoplanet Atmospheres with JWST. *AJ*, **158**(1), 27.

Lustig-Yaeger, J., V. S. Meadows, G. Tovar Mendoza, et al. (2018). Detecting Ocean Glint on Exoplanets Using Multiphase Mapping. *AJ*, **156**(6), p. 301.

Lyons, T. W., C. T. Reinhard, and N. J. Planavsky (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature*, **506**(7488), pp. 307–315.

MacDonald, M. G. (2019). Examining the radius valley: a machine-learning approach. *MNRAS*, **487**, pp. 5062–5069.

MacDonald, R. J. and N. Madhusudhan (2017). HD 209458b in new light: evidence of nitrogen chemistry, patchy clouds and sub-solar water. *MNRAS*, **469**, pp. 1979–1996.

Madden, J. H. and L. Kaltenegger (2018). A Catalog of Spectra, Albedos, and Colors of Solar System Bodies for Exoplanet Comparison. *Astrobiology*, **18**(12), pp. 1559–1573.

Madhusudhan, N. and A. Burrows (2012). Analytic Models for Albedos, Phase Curves, and Polarization of Reflected Light from Exoplanets. *ApJ*, **747**, 25.

Malin, M. C. and K. S. Edgett (2003). Evidence for Persistent Flow and Aqueous Sedimentation on Early Mars. *Science*, **302**(5652), pp. 1931–1934.

Mallama, A., B. Krobusek, and H. Pavlov (2017). Comprehensive wide-band magnitudes and albedos for the planets, with applications to exo-planets and Planet Nine. *Icarus*, **282**, pp. 19–33.

Mandel, K. and E. Agol (2002). Analytic Light Curves for Planetary Transit Searches. *ApJL*, **580**, pp. L171–L175.

Mann, H. B. and D. R. Whitney (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, **18**(1), pp. 50–60.

Marsh, T. R. (1989). The extraction of highly distorted spectra. *PASP*, **101**, pp. 1032–1037.

Martinez, C. F., K. Cunha, L. Ghezzi, and V. V. Smith (2019). A Spectroscopic Analysis of the California-Kepler Survey Sample. I. Stellar Parameters, Planetary Radii, and a Slope in the Radius Gap. *ApJ*, **875**, 29.

Mathur, S., T. S. Metcalfe, M. Woitaszek, et al. (2012). A Uniform Asteroseismic Analysis of 22 Solar-type Stars Observed by Kepler. *ApJ*, **749**(2), 152.

Maxted, P. F. L., C. Koen, and B. Smalley (2011). UBV(RI)$_C$ photometry of transiting planet hosting stars. *MNRAS*, **418**, pp. 1039–1042.

May, E. M., M. Zhao, M. Haidar, et al. (2018). MOPSS. I. Flat Optical Spectra for the Hot Jupiters WASP-4 b and WASP-52b. *AJ*, **156**, 122.

McCullough, P. R., N. Crouzet, D. Deming, and N. Madhusudhan (2014). Water Vapor in the Spectrum of the Extrasolar Planet HD 189733b. I. The Transit. *ApJ*, **791**, 55.

Meadows, V. S. (2017). Reflections on O_2 as a Biosignature in Exoplanetary Atmospheres. *Astrobiology*, **17**(10), pp. 1022–1052.

Meadows, V. S., G. N. Arney, E. W. Schwieterman, et al. (2018a). The habitability of proxima centauri b: environmental states and observational discriminants. *Astrobiology*, **18**(2), pp. 133–189.

Meadows, V. S., C. T. Reinhard, G. N. Arney, et al. (2018b). Exoplanet Biosignatures: Understanding Oxygen as a Biosignature in the Context of Its Environment. *Astrobiology*, **18**(6), pp. 630–662.

Meixner, M., A. Cooray, D. Leisawitz, et al. (2019). Origins Space Telescope Mission Concept Study Report. *arXiv e-prints*, arXiv:1912.06213.

Metchev, S. A., A. Heinze, D. Apai, et al. (2015). Weather on Other Worlds. II. Survey Results: Spots are Ubiquitous on L and T Dwarfs. *ApJ*, **799**, 154.

Milster, T., Y. Kim, Z. Wang, and K. Purvin (2020). Multiple-Order Diffractive Engineered Surface (MODE) Lenses. *Applied Optics*, **59**(26), pp. 7900–7906.

Mor, R., A. C. Robin, F. Figueras, et al. (2019). Gaia DR2 reveals a star formation burst in the disc 2-3 Gyr ago. *A&A*, **624**, L1.

Mulders, G. D., I. Pascucci, and D. Apai (2015a). A Stellar-mass-dependent Drop in Planet Occurrence Rates. *ApJ*, **798**(2), 112.

Mulders, G. D., I. Pascucci, and D. Apai (2015b). An Increase in the Mass of Planetary Systems around Lower-mass Stars. *ApJ*, **814**(2), 130.

Mulders, G. D., I. Pascucci, D. Apai, and F. J. Ciesla (2018). The Exoplanet Population Observation Simulator. I. The Inner Edges of Planetary Systems. *AJ*, **156**(1), p. 24.

Mullan, D. J. and H. P. Bais (2018). Photosynthesis on a Planet Orbiting an M Dwarf: Enhanced Effectiveness during Flares. *ApJ*, **865**(2), 101.

Murray, C. D. and A. C. M. Correia (2010). Keplerian Orbits and Dynamics of Exoplanets. In Seager, S. (ed.) *Exoplanets*, pp. 15–23.

Narita, N., T. Enomoto, S. Masaoka, and N. Kusakabe (2015). Titania may produce abiotic oxygen atmospheres on habitable exoplanets. *Scientific Reports*, **5**.

Neil, A. R. and L. A. Rogers (2020). A Joint Mass–Radius–Period Distribution of Exoplanets. *ApJ*, **891**(1), p. 12.

Nikolov, N., T. Henning, J. Koppenhoefer, et al. (2012). WASP-4b transit observations with GROND. *A&A*, **539**, A159.

Nikolov, N., D. K. Sing, A. S. Burrows, et al. (2015). HST hot-Jupiter transmission spectral survey: haze in the atmosphere of WASP-6b. *MNRAS*, **447**, pp. 463–478.

O'Neal, D., J. E. Neff, and S. H. Saar (1998). Measurements of Starspot Parameters on Active Stars using Molecular Bands in Echelle Spectra. *ApJ*, **507**, pp. 919–937.

O'Neal, D., J. E. Neff, S. H. Saar, and M. Cuntz (2004). Further Results of TiO-Band Observations of Starspots. *AJ*, **128**, pp. 1802–1811.

Owen, J. E. and S. Mohanty (2016). Habitability of terrestrial-mass planets in the HZ of M Dwarfs - I. H/He-dominated atmospheres. *MNRAS*, **459**, pp. 4088–4108.

Owen, J. E. and Y. Wu (2013). Kepler Planets: A Tale of Evaporation. *ApJ*, **775**, 105.

Owen, T. (1980). The Search for Early Forms of Life in Other Planetary Systems - Future Possibilities Afforded by Spectroscopic Techniques. In Papagiannis, M. D. (ed.) *Strategies for the Search for Life in the Universe*, volume 83, p. 177.

Parmentier, V., J. J. Fortney, A. P. Showman, et al. (2016). Transitions in the Cloud Composition of Hot Jupiters. *ApJ*, **828**, 22.

Pascucci, I., G. D. Mulders, and E. Lopez (2019). The Impact of Stripped Cores on the Frequency of Earth-size Planets in the Habitable Zone. *ApJ*, **883**(1), p. L15.

Pecaut, M. J. and E. E. Mamajek (2013). Intrinsic Colors, Temperatures, and Bolometric Corrections of Pre-main-sequence Stars. *ApJS*, **208**(1), 9.

Penny, M. T., B. S. Gaudi, E. Kerins, et al. (2019). Predictions of the WFIRST Microlensing Survey. I. Bound Planet Detection Rates. *ApJS*, **241**(1), 3.

Pidhorodetska, D., T. J. Fauchez, G. L. Villanueva, et al. (2020). Detectability of Molecular Signatures on TRAPPIST-1e through Transmission Spectroscopy Simulated for Future Space-based Observatories. *ApJL*, **898**(2), L33.

Pierrehumbert, R. and E. Gaidos (2011). Hydrogen Greenhouse Planets Beyond the Habitable Zone. *ApJL*, **734**(1), L13.

Pilcher, C. B. (2003). Biosignatures of Early Earths. *Astrobiology*, **3**(3), pp. 471–486.

Pinhas, A., B. V. Rackham, N. Madhusudhan, and D. Apai (2018). Retrieval of planetary and stellar properties in transmission spectroscopy with AURA. *MNRAS*, **480**(4), pp. 5314–5331.

Planavsky, N. J., D. Asael, A. Hofmann, et al. (2014). Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nature Geoscience*, **7**(4), pp. 283–286.

Plavchan, P., D. Latham, S. Gaudi, et al. (2015). Radial Velocity Prospects Current and Future: A White Paper Report prepared by the Study Analysis Group 8 for the Exoplanet Program Analysis Group (ExoPAG). *arXiv e-prints*, arXiv:1503.01770.

Pont, F., H. Knutson, R. L. Gilliland, et al. (2008). Detection of atmospheric haze on an extrasolar planet: the 0.55-1.05 μm transmission spectrum of HD 189733b with the HubbleSpaceTelescope. *MNRAS*, **385**, pp. 109–118.

Pont, F., D. K. Sing, N. P. Gibson, et al. (2013). The prevalence of dust on the exoplanet HD 189733b from Hubble and Spitzer observations. *MNRAS*, **432**, pp. 2917–2944.

Quanz, S. P., J. Kammerer, D. Defrère, et al. (2018). Exoplanet science with a space-based mid-infrared nulling interferometer. In *Proc. SPIE*, volume 10701 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 107011I.

Rackham, B., N. Espinoza, D. Apai, et al. (2017). ACCESS I: An Optical Transmission Spectrum of GJ 1214b Reveals a Heterogeneous Stellar Photosphere. *ApJ*, **834**, 151.

Rackham, B., A. Pinhas, D. Apai, et al. (2019a). Constraining Stellar Photospheres as an Essential Step for Transmission Spectroscopy of Small Exoplanets. *BAAS*, **51**(3), 328.

Rackham, B. V., D. Apai, and M. S. Giampapa (2018). The Transit Light Source Effect: False Spectral Features and Incorrect Densities for M-dwarf Transiting Planets. *ApJ*, **853**(2), 122.

Rackham, B. V., D. Apai, and M. S. Giampapa (2019b). The Transit Light Source Effect. II. The Impact of Stellar Heterogeneity on Transmission Spectra of Planets Orbiting Broadly Sun-like Stars. *AJ*, **157**(3), 96.

Ramirez, R. M. and L. Kaltenegger (2014). The Habitable Zones of Pre-main-sequence Stars. *ApJL*, **797**(2), L25.

Ramirez, R. M. and L. Kaltenegger (2017). A Volcanic Hydrogen Habitable Zone. *ApJL*, **837**(1), L4.

Ranjan, S., D. Charbonneau, J.-M. Désert, et al. (2014). Atmospheric Characterization of Five Hot Jupiters with the Wide Field Camera 3 on the Hubble Space Telescope. *ApJ*, **785**, 148.

Ranjan, S., R. Wordsworth, and D. D. Sasselov (2017). The Surface UV Environment on Planets Orbiting M Dwarfs: Implications for Prebiotic Chemistry and the Need for Experimental Follow-up. *ApJ*, **843**(2), 110.

Rauer, H., C. Catala, C. Aerts, et al. (2014). The PLATO 2.0 mission. *Experimental Astronomy*, **38**(1-2), pp. 249–330.

Reinhard, C. T., S. L. Olson, E. W. Schwieterman, and T. W. Lyons (2017). False Negatives for Remote Life Detection on Ocean-Bearing Planets: Lessons from the Early Earth. *Astrobiology*, **17**(4), p. 287.

Ribas, I., E. Bolmont, F. Selsis, et al. (2016). The habitability of Proxima Centauri b. I. Irradiation, rotation and volatile inventory from formation to the present. *A&A*, **596**, A111.

Ricker, G. R., J. N. Winn, R. Vanderspek, et al. (2014). Transiting Exoplanet Survey Satellite (TESS). In *Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave*, volume 9143 of *Proc. SPIE*, p. 914320.

Ricker, G. R., J. N. Winn, R. Vanderspek, et al. (2015). Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, **1**, 014003.

Rimmer, P. B., J. Xu, S. J. Thompson, et al. (2018). The origin of RNA precursors on exoplanets. *Science Advances*, **4**(8), p. eaar3302.

Ritchie, R. J., A. W. Larkum, and I. Ribas (2018). Could photosynthesis function on Proxima Centauri b? *International Journal of Astrobiology*, **17**(2), pp. 147–176.

Roberge, A., C. H. Chen, R. Millan-Gabet, et al. (2012). The Exozodiacal Dust Problem for Direct Observations of Exo-Earths. *PASP*, **124**, p. 799.

Robinson, T. D., V. S. Meadows, D. Crisp, et al. (2011). Earth as an Extrasolar Planet: Earth Model Validation Using EPOXI Earth Observations. *Astrobiology*, **11**, pp. 393–408.

Rodríguez, E., C. Rodríguez-López, M. J. López-González, et al. (2016). Search for pulsations in M dwarfs in the Kepler short-cadence data base. *MNRAS*, **457**(2), pp. 1851–1863.

Rodríguez-López, C., J. E. Gizis, J. MacDonald, et al. (2015). M dwarf search for pulsations within Kepler Guest Observer programme. *MNRAS*, **446**(3), pp. 2613–2620.

Rogers, L. A. (2015). Most 1.6 Earth-radius Planets are Not Rocky. *ApJ*, **801**, 41.

Rugheimer, S. and L. Kaltenegger (2018). Spectra of Earth-like Planets through Geological Evolution around FGKM Stars. *ApJ*, **854**(1), 19.

Sanchis-Ojeda, R., J. N. Winn, M. J. Holman, et al. (2011). Starspots and Spin-orbit Alignment in the WASP-4 Exoplanetary System. *ApJ*, **733**, 127.

Sanders, G. (2014). TMT Overview and Project Status. In Dickinson, M. and H. Inami (eds.) *Thirty Meter Telescope Science Forum*, p. 60.

Savransky, D. and D. Garrett (2016). WFIRST-AFTA coronagraph science yield modeling with EXOSIMS. *Journal of Astronomical Telescopes, Instruments, and Systems*, **2**, 011006.

Schwieterman, E. W., N. Y. Kiang, M. N. Parenteau, et al. (2018). Exoplanet Biosignatures: A Review of Remotely Detectable Signs of Life. *Astrobiology*, **18**(6), pp. 663–708.

Schwieterman, E. W., V. S. Meadows, S. D. Domagal-Goldman, et al. (2016). Identifying Planetary Biosignature Impostors: Spectral Features of CO and O_4 Resulting from Abiotic O_2/O_3 Production. *ApJL*, **819**, L13.

Seager, S., W. Bains, and R. Hu (2013). Biosignature gases in H2-Dominated atmospheres on rocky exoplanets. *ApJ*, **777**(2), p. 95.

Seager, S., W. Bains, and J. J. Petkowski (2016). Toward a List of Molecules as Potential Biosignature Gases for the Search for Life on Exoplanets and Applications to Terrestrial Biochemistry. *Astrobiology*, **16**, pp. 465–485.

Seager, S. and D. D. Sasselov (2000). Theoretical Transmission Spectra during Extrasolar Giant Planet Transits. *ApJ*, **537**, pp. 916–921.

Seager, S., B. A. Whitney, and D. D. Sasselov (2000). Photometric Light Curves and Polarization of Close-in Extrasolar Giant Planets. *ApJ*, **540**, pp. 504–520.

Segura, A., K. Krelove, J. F. Kasting, et al. (2003). Ozone concentrations and ultraviolet fluxes on earth-like planets around other stars. *Astrobiology*, **3**(4), pp. 689–708.

Shappee, B. J., J. L. Prieto, D. Grupe, et al. (2014). The Man behind the Curtain: X-Rays Drive the UV through NIR Variability in the 2013 Active Galactic Nucleus Outburst in NGC 2617. *ApJ*, **788**, 48.

Sharma, S., D. Stello, S. Buder, et al. (2018). The TESS-HERMES survey data release 1: high-resolution spectroscopy of the TESS southern continuous viewing zone. *MNRAS*, **473**(2), pp. 2004–2019.

Shields, A. L., S. Ballard, and J. A. Johnson (2016). The habitability of planets orbiting M-dwarf stars. *PhR*, **663**, p. 1.

Showman, A. P., J. J. Fortney, Y. Lian, et al. (2009). Atmospheric Circulation of Hot Jupiters: Coupled Radiative-Dynamical General Circulation Model Simulations of HD 189733b and HD 209458b. *ApJ*, **699**, pp. 564–584.

Silva Aguirre, V., G. R. Davies, S. Basu, et al. (2015). Ages and fundamental properties of Kepler exoplanet host stars from asteroseismology. *MNRAS*, **452**(2), pp. 2127–2148.

Sing, D. K., J. J. Fortney, N. Nikolov, et al. (2016). A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature*, **529**, pp. 59–62.

Sing, D. K., A. Lecavelier des Etangs, J. J. Fortney, et al. (2013). HST hot-Jupiter transmission spectral survey: evidence for aerosols and lack of TiO in the atmosphere of WASP-12b. *MNRAS*, **436**, pp. 2956–2973.

Sing, D. K., F. Pont, S. Aigrain, et al. (2011). Hubble Space Telescope transmission spectroscopy of the exoplanet HD 189733b: high-altitude atmospheric haze in the optical and near-ultraviolet with STIS. *MNRAS*, **416**, pp. 1443–1455.

Sing, D. K., H. R. Wakeford, A. P. Showman, et al. (2015). HST hot-Jupiter transmission spectral survey: detection of potassium in WASP-31b along with a cloud deck and Rayleigh scattering. *MNRAS*, **446**, pp. 2428–2443.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**(4), pp. 833–859.

Snaith, O., M. Haywood, P. Di Matteo, et al. (2015). Reconstructing the star formation history of the Milky Way disc(s) from chemical abundances. *A&A*, **578**, A87.

Snellen, I. A., R. J. De Kok, R. Le Poole, et al. (2013). Finding extraterrestrial life using ground-based high-dispersion spectroscopy. *ApJ*, **764**(2).

Solanki, S. K. (2003). Sunspots: An overview. *A&A Rv*, **11**, pp. 153–286.

Spake, J. J., D. K. Sing, T. M. Evans, et al. (2018). Helium in the eroding atmosphere of an exoplanet. *Nature*, **557**, pp. 68–70.

Speagle, J. S. (2020). DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *MNRAS*, **493**(3), pp. 3132–3158.

Staguhn, J., A. Mandell, K. Stevenson, et al. (2019). The Mid-InfraRed Exo-planet CLimate Explorer MIRECLE: Exploring the Nearest M-Earths Through Ultra-Stable Mid-IR Transit and Phase-Curve Spectroscopy. *BAAS*, **51**, p. 238.

Stahl, H. P. (2019). Multivariable Parametric Cost Model for Ground and Space Telescope Assemblies. *BAAS*, **51**, 143.

Stark, C. C., R. Belikov, M. R. Bolcar, et al. (2019). ExoEarth yield landscape for future direct imaging space telescopes. *Journal of Astronomical Telescopes, Instruments, and Systems*, **5**, 024009.

Stark, C. C., A. Roberge, A. Mandell, et al. (2015). Lower Limits on Aperture Size for an ExoEarth Detecting Coronagraphic Mission. *ApJ*, **808**(2), 149.

Stark, C. C., A. Roberge, A. Mandell, and T. D. Robinson (2014). Maximizing the ExoEarth candidate yield from a future direct imaging mission. *ApJ*, **795**(2), p. 122.

Stark, C. C., S. Shaklan, D. Lisman, et al. (2016). Maximized exoEarth candidate yields for starshades. *Journal of Astronomical Telescopes, Instruments, and Systems*, **2**(4), p. 041204.

Strassmeier, K. G. (2009). Starspots. *A&A Rv*, **17**, pp. 251–308.

Sudarsky, D., A. Burrows, and P. Pinto (2000). Albedo and Reflection Spectra of Extrasolar Giant Planets. *ApJ*, **538**, pp. 885–903.

Suissa, G., A. M. Mandell, E. T. Wolf, et al. (2020). Dim Prospects for Transmission Spectra of Ocean Earths around M Stars. *ApJ*, **891**(1), 58.

Sullivan, P. W., J. N. Winn, Z. K. Berta-Thompson, et al. (2015). The Transiting Exoplanet Survey Satellite: Simulations of Planet Detections and Astrophysical False Positives. *ApJ*, **809**(1), 77.

The LUVOIR Team (2019). The LUVOIR Mission Concept Study Final Report. *arXiv e-prints*, arXiv:1912.06219.

Tian, F., K. France, J. L. Linsky, et al. (2014). High stellar FUV/NUV ratio and oxygen contents in the atmospheres of potentially habitable planets. *Earth and Planetary Science Letters*.

Tian, F. and S. Ida (2015). Water contents of Earth-mass planets around M dwarfs. *Nature Geoscience*, **8**(3), pp. 177–180.

Traub, W. A., J. Breckinridge, T. P. Greene, et al. (2016). Science yield estimate with the Wide-Field Infrared Survey Telescope coronagraph. *Journal of Astronomical Telescopes, Instruments, and Systems*, **2**(1), 011020.

Traub, W. A. and B. R. Oppenheimer (2010). Direct Imaging of Exoplanets. In Seager, S. (ed.) *Exoplanets*, pp. 111–156.

Triaud, A. H. M. J., A. Collier Cameron, D. Queloz, et al. (2010). Spin-orbit angle measurements for six southern transiting planets. New insights into the dynamical origins of hot Jupiters. *A&A*, **524**, A25.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, **49**(2), pp. 71–104.

Tuomi, M., H. R. A. Jones, R. P. Butler, et al. (2019). Frequency of planets orbiting M dwarfs in the Solar neighbourhood. *arXiv e-prints*, arXiv:1906.04644.

Turbet, M., D. Ehrenreich, C. Lovis, et al. (2019). The runaway greenhouse radius inflation effect. An observational diagnostic to probe water on Earth-sized planets and test the habitable zone concept. *A&A*, **628**, A12.

Turbet, M., J. Leconte, F. Selsis, et al. (2016). The habitability of Proxima Centauri b. II. Possible climates and observability. *A&A*, **596**, A112.

Valencia, D., R. J. O'Connell, and D. D. Sasselov (2007). Inevitability of Plate Tectonics on Super-Earths. *ApJL*, **670**(1), pp. L45–L48.

Van Eylen, V. and S. Albrecht (2015). Eccentricity from Transit Photometry: Small Planets in Kepler Multi-planet Systems Have Low Eccentricities. *ApJ*, **808**(2), 126.

VanderPlas, J. T. (2018). Understanding the Lomb-Scargle Periodogram. *ApJS*, **236**, 16.

Villanueva, G. L., M. D. Smith, S. Protopapa, et al. (2018). Planetary Spectrum Generator: An accurate online radiative transfer suite for atmospheres, comets, small bodies and exoplanets. *JQSRT*, **217**, pp. 86–104.

Virtanen, P., R. Gommers, T. E. Oliphant, et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, pp. 261–272.

Wagner, K., A. Boehle, P. Pathak, et al. (2021). Imaging low-mass planets within the habitable zone of α Centauri. *Nature Communications*, **12**(1), pp. 1–7.

Wakeford, H. R., N. K. Lewis, J. Fowler, et al. (2019). Disentangling the Planet from the Star in Late-Type M Dwarfs: A Case Study of TRAPPIST-1g. *AJ*, **157**(1), 11.

Walker, J. C. G., P. B. Hays, and J. F. Kasting (1981). A negative feedback mechanism for the long-term stabilization of Earth's surface temperature. *Journal of Geophysical Research: Oceans*, **86**(C10), pp. 9776–9782.

Wall, J. V. and C. R. Jenkins (2003). *Practical statistics for astronomers*. Cambridge University Press.

Wang, J., D. Mawet, R. Hu, et al. (2018). Baseline requirements for detecting biosignatures with the HabEx and LUVOIR mission concepts. *Journal of Astronomical Telescopes, Instruments, and Systems*, **4**, 035001.

Way, M. J., A. D. Del Genio, N. Y. Kiang, et al. (2016). Was Venus the first habitable world of our solar system? *Geophysical Research Letters*, **43**(16), pp. 8376–8383.

Weinberger, A., J. Berkby, B. Bowler, et al. (2018). Exoplanets & Planet Formation. In *GMT Science Book 2018*, pp. 31–50. Https://www.gmto.org/gallery/gmt-resources/.

Weiss, L. M. and G. W. Marcy (2014). The Mass-Radius Relation for 65 Exoplanets Smaller than 4 Earth Radii. *ApJL*, **783**, L6.

Wilson, D. M., M. Gillon, C. Hellier, et al. (2008). WASP-4b: A 12th Magnitude Transiting Hot Jupiter in the Southern Hemisphere. *ApJL*, **675**, L113.

Winn, J. N. (2010). Transits and Occultations. In Seager, S. (ed.) *Exoplanets*, pp. 55–77.

Winn, J. N., M. J. Holman, J. A. Carter, et al. (2009). The Transit Light Curve Project. XI. Submillimagnitude Photometry of Two Transits of the Bloated Planet WASP-4b. *AJ*, **137**, pp. 3826–3833.

Wolfgang, A., L. A. Rogers, and E. B. Ford (2016). Probabilistic Mass-Radius Relationship for Sub-Neptune-Sized Planets. *ApJ*, **825**(1), 19.

Wordsworth, R. and R. Pierrehumbert (2014). Abiotic oxygen-dominated atmospheres on terrestrial habitable zone planets. *ApJL*, **785**(2).

Wordsworth, R. D., L. K. Schaefer, and R. A. Fischer (2018). Redox Evolution via Gravitational Differentiation on Low-mass Planets: Implications for Abiotic Oxygen, Water Loss, and Habitability. *AJ*, **155**(5), 195.

Yang, J., G. Boué, D. C. Fabrycky, and D. S. Abbot (2014). Strong Dependence of the Inner Edge of the Habitable Zone on Planetary Rotation Rate. *ApJL*, **787**(1), L2.

Zahnle, K., M. Claire, and D. Catling (2006). The loss of mass-independent fractionation in sulfur due to a Palaeoproterozoic collapse of atmospheric methane. *Geobiology*, **4**(4), pp. 271–283.

Zahnle, K. J. and D. C. Catling (2013). The Cosmic Shoreline. In *Lunar and Planetary Science Conference*, volume 44 of *Lunar and Planetary Inst. Technical Report*, p. 2787.

Zahnle, K. J. and D. C. Catling (2017). The Cosmic Shoreline: The Evidence that Escape Determines which Planets Have Atmospheres, and what this May Mean for Proxima Centauri B. *ApJ*, **843**(2), p. 122.

Zahnle, K. J., D. C. Catling, and M. W. Claire (2013). The rise of oxygen and the hydrogen hourglass. *Chemical Geology*, **362**, pp. 26–34.

Zahnle, K. J., M. Gacesa, and D. C. Catling (2019). Strange messenger: A new history of hydrogen on Earth, as told by Xenon. *Geochimica et Cosmochimica Acta*, **244**, pp. 56–85.

Zhang, Z., Y. Zhou, B. V. Rackham, and D. Apai (2018). The Near-infrared Transmission Spectra of TRAPPIST-1 Planets b, c, d, e, f, and g and Stellar Contamination in Multi-epoch Transit Spectra. *AJ*, **156**(4), 178.

Zhou, Y., D. Apai, G. H. Schneider, et al. (2016). Discovery of Rotational Modulations in the Planetary-mass Companion 2M1207b: Intermediate Rotation Period and Heterogeneous Clouds in a Low Gravity Atmosphere. *ApJ*, **818**, 176.

Zsom, A., S. Seager, J. de Wit, and V. Stamenković (2013). Toward the Minimum Inner Edge Distance of the Habitable Zone. *ApJ*, **778**(2), 109.