Characterizing Earth Analogs in Reflected Light: Atmospheric Retrieval Studies for Future Space Telescopes

Y. Katherina Feng[1,2,11] , Tyler D. Robinson[1,3,4] , Jonathan J. Fortney[1,2] , Roxana E. Lupu[5,6] , Mark S. Marley[5] ,
Nikole K. Lewis[7,8] , Bruce Macintosh[9] , and Michael R. Line[10]

[1] Department of Astronomy & Astrophysics, University of California, Santa Cruz, CA 95064, USA
[2] University of California, Santa Cruz, Other Worlds Laboratory, USA
[3] Now at Department of Physics & Astronomy, Northern Arizona University, Flagstaff, AZ 86011, USA
[4] NASA Astrobiology Institute's Virtual Planetary Laboratory, University of Washington, Seattle, WA 98195, USA
[5] NASA Ames Research Center, Moffett Field, CA 94035, USA
[6] Bay Area Environmental Research Institute, Petaluma, CA 94952, USA
[7] Space Telescope Science Institute, Baltimore, MD 21218, USA
[8] Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218, USA
[9] Department of Physics, Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA
[10] School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287, USA

Abstract

Space-based high-contrast imaging mission concepts for studying rocky exoplanets in reflected light are currently under community study. We develop an inverse modeling framework to estimate the science return of such missions given different instrument design considerations. By combining an exoplanet albedo model, instrument noise model, and ensemble Markov chain Monte Carlo sampler, we explore retrievals of atmospheric and planetary properties for Earth twins as a function of signal-to-noise ratio (S/N) and resolution (R). Our forward model includes Rayleigh-scattering, single-layer water clouds with patchy coverage, and pressure-dependent absorption due to water vapor, oxygen, and ozone. We simulate data at $R = 70$ and 140 from 0.4 to 1.0 μm with S/N = 5, 10, 15, and 20 at 550 nm (i.e., for *HabEx*/*LUVOIR*-type instruments). At these same S/Ns, we simulate data for *WFIRST* paired with a starshade, which includes two photometric points between 0.48 and 0.6 μm and $R = 50$ spectroscopy from 0.6 to 0.97 μm. Given our noise model for *WFIRST*-type detectors, we find that weak detections of water vapor, ozone, and oxygen can be achieved with observations with at least $R = 70/\mathrm{S/N} = 15$ or $R = 140/\mathrm{S/N} = 10$ for improved detections. Meaningful constraints are only achieved with $R = 140/\mathrm{S/N} = 20$ data. The *WFIRST* data offer limited diagnostic information, needing at least S/N = 20 to weakly detect gases. Most scenarios place limits on planetary radius but cannot constrain surface gravity and, thus, planetary mass.

Key words: methods: statistical – planets and satellites: atmospheres – planets and satellites: terrestrial planets

1. Introduction

The scientific field of exoplanets has been rapidly advancing since the hallmark discovery of the first planet orbiting a Sun-like star (Mayor & Queloz 1995). Following the launch of NASA's *Kepler* mission (Borucki et al. 2003, 2011), the field has seen the discovery of thousands of transiting exoplanets and the exciting result that planets with radii between 0.75 and 2.5 $R_\oplus$ are common around solar-type stars (Burke et al. 2015). Only within the last decade have observational studies for exoplanet atmospheric characterization seen substantial development, starting with the first detection of an exoplanet's atmosphere by Charbonneau et al. (2002).

To date, the majority of exoplanet atmospheric characterization investigations have focused on transiting worlds. Hot Jupiters, owing to their large sizes and short orbital periods, are typically emphasized as targets for these studies. Characterization of small, potentially rocky exoplanets is limited to worlds with cool stellar hosts (K and M dwarfs), which offer favorable planet-to-star size ratios. Recently, de Wit et al. (2016) studied the combined transmission spectra of two transiting Earth-sized planets orbiting the ultracool dwarf TRAPPIST-1 using the *Hubble Space Telescope*. While no gas absorption features were detected by de Wit et al. (2016), this work highlights the

improvements in signal size when terrestrial-sized transiting planets are studied around low-mass stars. Additionally, since the habitable zone (Kasting et al. 1993) around a low-mass star is relatively close-in, characterization studies of potentially habitable exoplanets around cool stars can benefit from the frequency of transit events. However, for Sun-like hosts, the planet-to-star size ratio is much less favorable, and the habitable zone is located far from the star, thus severely limiting the potential for atmospheric characterization.

Direct, high-contrast imaging has now emerged as an essential technique for studying the atmospheres of planets at larger orbital separations from their host star (i.e., at orbital distances $\gtrsim$1 au). Thus far, high-contrast imaging has been proven successful in studying atmospheres of young, self-luminous gas giants in the near- and mid-infrared (e.g., Barman et al. 2011; Skemer et al. 2014; Macintosh et al. 2015). These worlds, owing to their intrinsic brightness, have typical contrast ratios of 10^{-4} with respect to their hosts. A true Jupiter analog at visible wavelengths, by comparison, would have a contrast ratio of 10^{-9}, while an Earth analog would have a contrast ratio of order 10^{-10}. Reflected light in the visible probes to atmospheric depths of up to ~10 bar for giant planets (Marley et al. 2014), which is complimentary to the relatively low pressures probed in transit observations (typically less than 10–100 mbar). Additionally, the wavelength range of 0.4–1.0 μm holds rich information about a planet's atmosphere,

---

including signatures of methane, water vapor, and haze (Burrows 2014; Marley et al. 2014).

In spite of the incredible technological challenges, there are multiple planned or in-development space-based missions that would be capable of high-contrast imaging of exoplanets in reflected light. First among these will be NASA's *Wide-Field InfraRed Survey Telescope* (*WFIRST*; Spergel et al. 2013), which was identified as the top priority space mission in the 2010 National Academy of Sciences Decadal Survey of Astronomy and Astrophysics.[12] The *WFIRST* mission will carry a coronagraphic instrument (CGI) with imaging capability and a visible-light integral field spectrograph of wavelength resolution ∼50 (Balasubramanian et al. 2016; Cady et al. 2016; Noecker et al. 2016; Seo et al. 2016; Trauger et al. 2016; Groff et al. 2018). Although envisioned primarily as a technology demonstrator, it may study the atmospheres of relatively cool gas giant exoplanets that have been previously detected using the radial velocity technique (Traub et al. 2016).

While *WFIRST* could also have some capability to survey stars in the solar neighborhood for lower-mass planetary companions (Burrows 2014; Greco & Burrows 2015; Spergel et al. 2015; Robinson et al. 2016; Savransky & Garrett 2016), it is anticipated that the core optical throughput of the *WFIRST* CGI will be low for planetary signals. This stems primarily from the complexities of accommodating for *WFIRST*'s on-axis secondary mirror and support structures within the high-contrast instruments (Krist et al. 2016; Traub et al. 2016). Low throughput drives long requisite integration times, thereby likely making spectroscopic observations of smaller, less-bright worlds (such as super-Earth exoplanets) unfeasible except around the very closest stars (Robinson et al. 2016). However, if the *WFIRST* spacecraft were to be paired with an external starshade (Cash 2006; Kasdin et al. 2012), the CGI can be operated in a direct mode without coronagraphic masks, substantially increasing throughput. High-contrast imaging of sub-Neptune and terrestrial-sized exoplanets may then become possible. The feasibility of a starshade "rendezvous" with the *WFIRST* spacecraft is under active investigation (Seager et al. 2015; Crill & Siegler 2017).

In advance of the 2020 astronomy and astrophysics decadal survey, several large-scale space-based mission concepts are being studied.[13] Of these, two have a strong focus on the characterization of rocky exoplanets with direct imaging: the *Habitable Exoplanet Imaging Mission* (*HabEx*; Mennesson et al. 2016) and the *Large Ultra-Violet/Optical/InfraRed Surveyor* (*LUVOIR*; Peterson et al. 2017). *HabEx* and *LUVOIR* are incorporating aspects of design that would allow the detection of water vapor and biosignatures on planets in the habitable zones of nearby Sun-like stars. It is therefore timely and critical that we explore observational approaches that maximize science yield during the development of these large-scale mission concepts, as well as the *WFIRST* rendezvous concept. To accomplish this, we must perform atmospheric and instrument modeling to simulate the types of spectra we can expect to measure, and we must develop tools to infer planetary properties from these simulated observations.

Traditionally, the comparison to a limited range of forward models has been used to infer atmospheric properties (such as temperature structure and gas abundances) from spectral

observations. This involves iterating to a radiative-convective solution for a given set of planetary parameters (e.g., gravity, metallicity, equilibrium abundances, incident flux), and can include detailed treatment of aerosols, chemistry, and dynamics within the model atmosphere (Marley & Robinson 2015). The goal is to generate a spectrum that matches the available data and thus offers one potential explanation for the world's atmospheric state (e.g., Konopacky et al. 2013; Barman et al. 2015; Macintosh et al. 2015). A more data-driven interpretation of atmospheric observations is accomplished through inverse modeling, or retrievals. Developed for solar system studies and remote sensing (e.g., Rodgers 1976; Irwin et al. 2008), retrievals have become a valuable tool in constraining our understanding of the atmospheres of transiting exoplanets. Early exoplanet retrieval work invoked grid-based optimization schemes (Madhusudhan & Seager 2009), while subsequent works have taken advantage of Bayesian inference with methods such as optimal estimation and Markov chain Monte Carlo (MCMC; e.g., Benneke & Seager 2012; Lee et al. 2012; Line et al. 2013).

Several studies have examined the hypothetical yield from characterizing giant exoplanets observed with a space-based coronagraph (such as *WFIRST*) with retrieval techniques. Marley et al. (2014), for example, modeled spectra we could expect from known radial velocity gas giants if observed by the *WFIRST* CGI. Given the diversity of cool giant planets, the model spectra have a variety of input assumptions for clouds, surface gravity, and atmospheric metallicity. Marley et al. (2014) then applied retrieval methods to these synthetic spectra, enabling the exploration of how well atmospheric parameters are constrained under varying quality of data. Lupu et al. (2016) further investigated the feasibility of characterizing cool giant planet atmospheres through retrieval, focusing on the ability to constrain the CH_4 abundance and cloud properties. The systematic study of the impact of conditions like signal-to-noise ratios (S/Ns) or wavelength resolution is essential to quantifying the scientific return of these reflected-light observations. Nayak et al. (2017) considered the impact of an unknown phase angle on the inference of properties such as planet radius and gravity. In all of these studies, the S/N of the data has a significant influence on the constraints of atmospheric properties.

Previous work on smaller planets in the context of possible future space missions includes von Paris et al. (2013), who synthesized infrared emission observations of a cloud-free, directly imaged Earth twin and employed a least-squares approach and χ^2 maps to perform retrievals and explore parameter space (considering the effects of instrument resolution and S/Ns). A collection of recent studies (Mawet et al. 2017; Wang et al. 2017a, 2017b) examined atmospheric species detection using "high-dispersion coronagraphy," which couples starlight-suppression technologies with high-resolution spectroscopy. In these studies, simulated observations (typically at spectral resolutions, $R = \lambda/\Delta\lambda$, of many hundreds to tens of thousands) are cross-correlated with template molecular opacity spectra to explore the feasibility of species detection. While this novel approach can yield detections of key atmospheric constituents, the abundance of these atmospheric species cannot be robustly constrained.

To date, there still does not exist a systematic study of the atmospheric characterization of small exoplanets using retrieval techniques on reflected-light observations at spectral

---

[12] http://sites.nationalacademies.org/bpa/bpa_049810
[13] https://science.nasa.gov/astrophysics/2020-decadal-survey-planning

resolutions relevant to *WFIRST* rendezvous, *HabEx*, and *LUVOIR*. Motivated by this need, we present here our extension of Bayesian retrieval techniques into the terrestrial regime. We construct a forward model suitable for simulating reflectance spectra of Earth-like planets in the visible wavelength range of 0.4 to 1.0 μm. We explore retrievals of planetary and atmospheric properties from simulated data sets at varying spectral resolutions and S/Ns. A retrieval framework such as this allows us to quantify uncertainties we expect for key planetary parameters given certain observing scenarios. Thus, our approach enables us to search for the minimal observing conditions that achieve the scientific goal of identifying traits associated with habitability and life. In particular, we are interested in our ability to detect and constrain abundances of molecules such as water, oxygen, and ozone; characterize basic properties of a cloud layer; and measure bulk parameters such as radius.

In Section 2, we describe our forward model and construction of simulated data. In Section 3, we validate our forward model by building up retrieval complexity (i.e., number of retrieved parameters). We perform a study of retrieval performance with respect to spectral resolution and S/N in Section 4, with implications for *HabEx*/*LUVOIR*. We also study the retrieval performance for data sets expected from a *WFIRST* rendezvous scenario, where the CGI would provide modest-resolution spectroscopy in the red (600–970 nm) and photometry in the blue (480–600 nm). We present our discussion and conclusions in Sections 5 and 6, respectively.

2. Methods

The observed quantity for a directly imaged exoplanet in reflected light at a given phase (i.e., planet–star–observer) angle, α, is the wavelength-dependent planet-to-star flux ratio,

$$\frac{F_p}{F_s} = A_g \Phi(\alpha) \left(\frac{R_p}{r} \right)^2, \qquad (1)$$

where A_g is the geometric albedo, $\Phi(\alpha)$ is the phase function, R_p is the radius of the planet, and r is the orbital separation. The phase function (which depends on wavelength) translates the planetary brightness at full phase (i.e., where $\alpha = 0°$) to its brightness at different phase angles. The wavelength-dependent geometric albedo is the ratio of the measured flux from the planet at full phase to that from a perfectly reflecting Lambert (i.e., isotropically reflecting) disk with the size of the planet. We denote the product of the geometric albedo and the phase function as the phase-dependent "reflectance" of the planet. In general, the geometric albedo encodes information about the composition and structure (i.e., "state") of an atmosphere, while the phase function is strongly related to the scattering properties of an atmosphere (e.g., Marley et al. 1999; Burrows 2014).

To understand the information contained in direct-imaging observations of exoplanets in reflected light, we employ a retrieval (or inverse analysis) framework that consists of several linked simulation tools and models. Of central importance is a well-tested three-dimensional albedo model—described in greater depth below—that computes a reflectance spectrum at high resolution for a planet given a description of its atmospheric state (McKay et al. 1989; Marley et al. 1999; Cahoy et al. 2010; Lupu et al. 2016; Nayak et al. 2017). When

coupled with a simulator for degrading a high-resolution spectrum to match the resolution of an instrument, we refer to these two tools as the "forward model." By adding simulated noise to forward-model spectra, we generate faux "observations" of worlds as would be studied by future high-contrast imaging missions. To create "observed" spectra, we adopt a widely used direct-imaging instrument simulator (Robinson et al. 2016) that generates synthetic observations given an input, noise-free spectrum.

Given an "observed" planet-to-star flux ratio spectrum, our inverse analyses use a Bayesian inference tool that compares the observation to forward-model outputs to sample the posterior probability distributions for a collection of atmospheric-state parameters. In other words, our inverse analyses indicate what range of atmospheric-state parameters (e.g., gas abundances) adequately describe a direct-imaging observation. Our Bayesian parameter estimations use an open-source affine-invariant MCMC ensemble sampler, emcee (Goodman & Weare 2010; Foreman-Mackey et al. 2013).

In this work, retrieval analyses generally proceed by first simulating a noise-free spectrum of a world with a known atmospheric state (e.g., Earth). We then add simulated observational noise to this spectrum to create a synthetic observation. Following Bayesian parameter estimation on this synthetic observation, we can compare a retrieved atmospheric state to the original, known atmospheric state, thereby allowing us to understand how observational noise affects our ability to deduce the true nature of an exoplanetary atmosphere.

2.1. Albedo Model

Our three-dimensional albedo model (see also Cahoy et al. 2010) divides a world into a number of plane-parallel facets with coordinates of longitude (ζ) and colatitude (η), with the former referenced from the subobserver location and the latter ranging from zero at the northern pole to π at the southern pole. A single facet has downwelling incident stellar radiation from a zenith angle $\mu_s = \cos \theta_s = \sin \eta \cos(\zeta - \alpha)$, where, as earlier, α is the phase angle. The facet reflects to the observer in a direction whose zenith angle is given by $\mu_o = \cos \theta_o = \sin \eta \cos \zeta$. Note that, at full phase (where the geometric albedo is defined), the observer and source are colinear such that $\mu_o = \mu_s$ for all facets.

The atmosphere above each facet is divided into a set of pressure levels, and we perform a radiative transfer calculation to determine the emergent intensity. With the intensities calculated for an entire visible hemisphere, we follow the methods outlined by Horak (1950) and Horak & Little (1965) to perform integration using Chebychev–Gauss quadrature, thus producing the reflectance value at a given wavelength. We repeat this procedure at each of the wavelength points within a specified range to build up a reflectance spectrum.

Taking $I(\tau, \mu, \phi)$ to be the wavelength-dependent intensity at optical depth τ in a direction determined by the zenith and azimuth angles μ and ϕ, we ultimately need to determine the emergent intensity from each facet in the direction of the observer, $I(\tau = 0, \mu_o, \phi_o)$. Thus, for each facet, we must solve the one-dimensional, plane-parallel radiative transfer equation,

$$\mu \frac{dI}{d\tau} = I(\tau, \mu, \phi) - S(\tau, \mu, \phi), \qquad (2)$$

where S is the wavelength-dependent source function. Following Meador & Weaver (1980), Toon et al. (1989), and

Marley & Robinson (2015), the source function is

$$S(\tau, \mu, \phi) = \frac{\bar{\omega}}{4\pi} F_s \cdot p(\tau, \mu, \phi, -\mu_s, \phi_s) \cdot e^{-\tau/\mu_\odot}$$
$$+ \bar{\omega} \int_0^{2\pi} d\phi' \int_{-1}^1 \frac{d\mu'}{4\pi} \cdot I(\tau, \mu', \phi') \cdot p(\tau, \mu, \phi, \mu', \phi'),$$
(3)

where $\bar{\omega}$ is the single-scattering albedo, F_s is the incoming stellar flux at the top of the atmosphere (which we normalize to unity so that emergent intensities correspond to reflectivities), ϕ_s is the stellar azimuth angle, and p is the scattering phase function. Note that our source function does not include an emission term, since we are not computing thermal spectra. Recall that the first term in Equation (3) describes directly scattered radiation from the direct solar beam, while the final term describes diffusely scattered radiation from the (μ', ϕ') direction scattering into the (μ, ϕ) direction.

Like most standard tools for solving the radiative transfer equation, we separate treatments of directly scattered radiation from diffusely scattered radiation, and, for both, it is convenient to express the scattering phase function in terms of a unique scattering angle, Θ. As single-scattered radiation typically has more distinct forward- and backward-scattering features, we choose to represent the scattering phase function for the direct beam with a two-term Henyey–Greenstein (TTHG) phase function (Kattawar 1975),

$$p_{\mathrm{TTHG}}(\Theta) = f p_{\mathrm{HG}}(g_f, \Theta) + (1 - f) p_{\mathrm{HG}}(g_b, \Theta),$$
(4)

where p_{HG} is the Henyey–Greenstein (HG) phase function with

$$p_{\mathrm{HG}} = \frac{1}{4\pi} \frac{1 - \bar{g}^2}{(1 + \bar{g}^2 - 2\bar{g}\cos\Theta)^{3/2}}.$$
(5)

Recall that a TTHG phase function can represent both forward- and backward-scattering peaks, while the (one-term) HG phase function only has one peak (typically in the forward direction). In the previous expressions, $\bar{g}$ is the asymmetry parameter, f is the forward/backward scatter fraction, g_f is the asymmetry parameter for the forward-scattered portion of the TTHG, and g_b is the asymmetry parameter for the backward-scattered portion of the TTHG. For the forward- and backward-scattering portions of the TTHG phase function, we use $g_f = \bar{g}$, $g_b = -\bar{g}/2$, and $f = 1 - g_b^2$. Substituting these into the TTHG phase function expression yields

$$p_{\mathrm{TTHG}} = \frac{\bar{g}^2}{4} p_{\mathrm{HG}}(-\bar{g}/2, \Theta) + \left(1 - \frac{\bar{g}^2}{4}\right) p_{\mathrm{HG}}(\bar{g}, \Theta).$$
(6)

For radiation that is single-scattered from the solar beam to the observer, the scattering geometry is fixed by the planetary phase angle such that $\Theta = \pi - \alpha$. Our choice of parameters, and their relation to $\bar{g}$, in the TTHG was designed by Cahoy et al. (2010) to roughly reproduce the phase function of liquid-water clouds. This parameterization, however, is different from that proposed by Kattawar (1975). We do not expect our results to be sensitive to the details of a particular phase function treatment, as Lupu et al. (2016) showed that scattered-light retrievals struggle to constrain phase function parameters.

We adopt a standard two-stream approach to solving the radiative transfer equation (Meador & Weaver 1980). In this case, the diffusely scattered component of the source function is azimuthally averaged. Combined with our representation of the directly scattered component, we have

$$S(\tau, \mu, \mu_s, \alpha) = \frac{\bar{\omega}}{4\pi} F_s \cdot p_{\mathrm{TTHG}}(\mu, -\mu_s) \cdot e^{-\tau/\mu_s}$$
$$+ \frac{\bar{\omega}}{4\pi} \int_{-1}^1 I(\tau, \mu') p(\mu, \mu') d\mu',$$
(7)

where the azimuth-averaged phase functions are given by

$$p(\mu, \mu') = \frac{1}{2\pi} \int_0^{2\pi} p(\mu, \phi, \mu', \phi) d\phi.$$
(8)

We represent the azimuth-averaged scattering phase functions as a series of Legendre polynomials, $P_l(\mu)$, expanded to order M with

$$p(\mu, \mu') = \sum_{l=0}^M g_l P_l(\mu) P_l(\mu'),$$
(9)

where the phase function moments, g_l, are defined according to

$$g_l = \frac{2l + 1}{2} \int_{-1}^1 p(\cos\Theta) P_l(\cos\Theta) d\cos\Theta.$$
(10)

The first moment of the phase function is related to the asymmetry parameter, with $\bar{g} = g_1/3$. We use a second-order expansion of the phase function, giving

$$p(\mu, \mu') = 1 + 3\bar{g}\mu\mu' + \frac{g_2}{2}(3\mu^2 - 1)(3\mu'^2 - 1).$$
(11)

In a given atmospheric layer of our albedo model, the optical depth is the sum of the scattering optical depth and the absorption optical depth, $\tau = \tau_{\mathrm{scat}} + \tau_{\mathrm{abs}}$. The scattering optical depth has contributions from Rayleigh scattering and clouds, so that $\tau_{\mathrm{scat}} = \tau_{\mathrm{Ray}} + \bar{\omega}_{\mathrm{cld}}\tau_{\mathrm{cld}}$, where $\bar{\omega}_{\mathrm{cld}}$ is the cloud single-scattering albedo. The single-scattering albedo for a layer is then $\bar{\omega} = \tau_{\mathrm{scat}}/\tau$. We determine the asymmetry parameter, $\bar{g}$, with an optical depth weighting on the Rayleigh-scattering asymmetry parameter (which is zero) and the cloud-scattering asymmetry parameter, yielding $\bar{g} = \bar{g}_{\mathrm{cld}}(\tau_{\mathrm{cld}}/\tau_{\mathrm{scat}})$. When representing the second moment of the phase function, we use $g_2 = \frac{1}{2}(\tau_{\mathrm{Ray}}/\tau_{\mathrm{scat}})$ so that g_2 tends toward the appropriate value for Rayleigh scattering (i.e., $1/2$; Hansen & Travis 1974) when the Rayleigh-scattering optical depth dominates the scattering optical depth.

2.2. Model Upgrades

As compared to prior investigations that have used the Cahoy et al. (2010) albedo tool (e.g., Lupu et al. 2016; Nayak et al. 2017), we have updated the model to include an optional isotropically reflecting (Lambertian) lower boundary (mimicking a planetary surface) and added pressure-dependent absorption due to H_2O, O_3, O_2, and CO_2. As in previous studies, CH_4 remains a radiatively active species in the model. We also include Rayleigh scattering from H_2O, O_2, CO_2, and N_2 (in addition to H_2 and He from previous studies). As in Lupu et al. (2016), we allow for an extended gray-scattering cloud in our atmospheres.

In Lupu et al. (2016), their two-layer cloud model atmosphere includes a deeper, optically thick cloud deck that
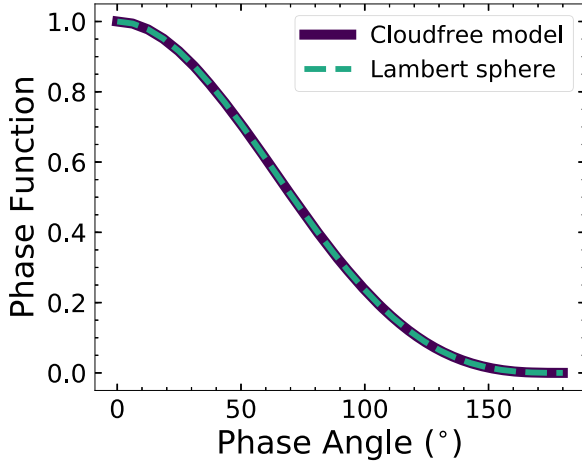
Figure 1. Comparing our model phase function to the analytic Lambertian phase function (Equation (12)). No atmospheric absorption or scattering is present in the forward model.
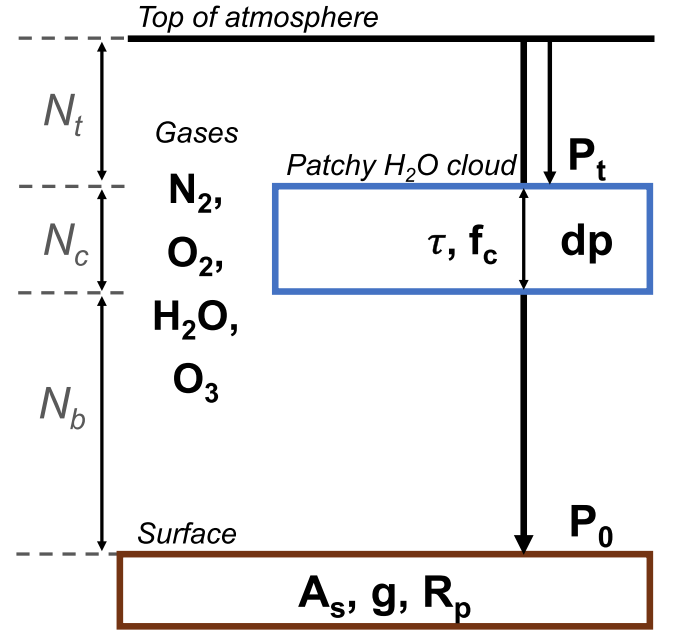

Figure 2. Illustrative schematic of our model atmosphere's structure. The atmosphere has $N_t + N_c + N_b$ layers. Table 1 lists the definitions, fiducial values, and priors of the presented parameters.

essentially acts as a reflective surface. Unlike the gas giants within that study, though, terrestrial planets have a solid surface we can probe. We characterize our isotropically reflecting lower boundary using a spherical albedo for the planetary surface, A_s, which represents the specific power in scattered, outgoing radiation compared to that in incident radiation. For this study, we simply adopt gray surface albedo values, which reduces complexity and computation time. For the inhomogeneous surface of a realistic Earth, featuring oceans and continents, the surface albedo is wavelength-dependent, and we hope to investigate the significance of such surfaces in future work.

We undertook a test to check our reflective lower-boundary condition in the limit of a transparent atmosphere. Without atmospheric absorption or scattering, our assumption of a Lambertian surface would imply that the reflectivity (or phase function) determined by our albedo code should follow the analytic Lambert phase function,

$$\Phi_L(\alpha) = \frac{\sin\alpha + (\pi - \alpha)\cos\alpha}{\pi}. \quad (12)$$

Figure 1 compares the model phase function with the analytic phase function and shows complete agreement, confirming that our treatment of the surface is correct.

Previous work featuring the albedo model adopted here used a predefined atmospheric pressure grid. To accommodate the finite surface pressures of rocky planets, as well as the various combinations of cloud parameters our retrievals will explore, we instead establish an adaptive method of determining the pressure grid. Here we divide the atmosphere into a pressure grid of N_{level}, bounded by $P = P_{top}$ at the top of the atmosphere and $P = P_0$ at the surface. In a cloud-free scenario, we simply divide the atmosphere evenly in log-P space.

For our simulations that include a single cloud deck, we adaptively determine the pressure value at each level depending on the location, thickness, and optical depth of the cloud. The quantities that define the cloud deck are p_t, the cloud-top pressure; dp, the atmospheric pressure across the cloud; and τ, the cloud optical depth. We begin by assigning a number of layers to the cloud, imposing two conditions: (1) there should be at least three model pressure layers to each atmospheric pressure scale height (perH = 3), and (2) the cloud optical

depth in a layer must remain below at most 5 (maxtau = 5). This allows us to avoid any one layer spanning a large extent within the atmosphere and also avoids cloud layers that have extremely large scattering optical depths.

When beginning our gridding process, we propose an initial number of cloud layers, $N_c = $ perH $\times$ numH, where numH $= \ln\frac{p_t + dp}{p_t}$ is the number of e-folding distances through the cloud (serving as a proxy for scale height). The aerosol optical depth for each pressure layer within the cloud would then simply be $\Delta\tau = \frac{\tau}{N_c}$. However, if $\Delta\tau > $ maxtau, we adjust the cloud resolution by increasing N_c by a factor of $\frac{\Delta\tau}{\text{maxtau}}$ and then round up to the nearest integer. In other words, we increase the resolution of the pressure grid through the cloud until the layer optical depth is under maxtau. We determine successive pressure level values through the cloud with $p[i] = p[i-1] + \Delta\ln p$, where $\Delta\ln p = \frac{\ln(p_t + dp) - \ln p_t}{N_c}$, starting from the top of the cloud. We divide the remaining $N_{level} - N_c$ levels in uniform $\ln p$ space on either side of the cloud, weighted by the number of pressure scale heights above (N_t) and below (N_b) the cloud. Figure 2 visualizes the three portions of the atmosphere.

For simplicity, we assume an isothermal atmosphere (at $T = 250$ K), as temperature has little effect on the reflected-light spectrum (Robinson 2017). Pressure, however, has a strong impact on molecular opacities, as seen in Figure 3. We incorporated high-resolution pressure-dependent opacities for all molecules in our atmosphere. The absorption opacities are generated line-by-line from the HITRAN2012 line list (Rothman et al. 2013) for seven orders of magnitude in pressure (10^{-5}–10^2 bar) at T, spanning our entire wavelength range at <1 cm^{-1} resolution. Figure 3 also illustrates how absorption features of H_2O, O_2, and O_3 change when in an atmosphere of 1 bar versus one of 10 bars.

We interpolate our high-resolution opacity tables to the slightly lower resolution of the forward model in order to maintain short model runtimes while not affecting the accuracy
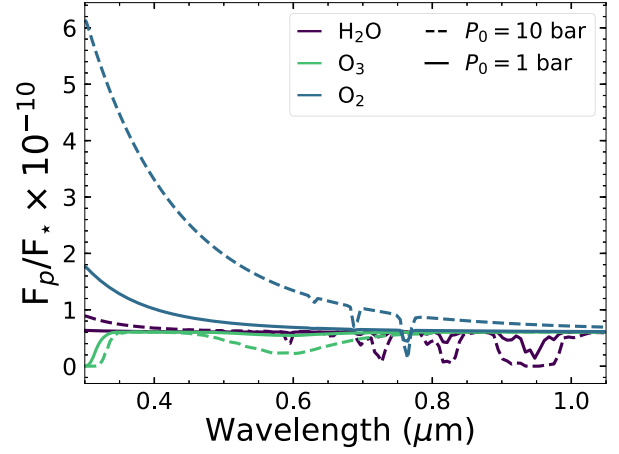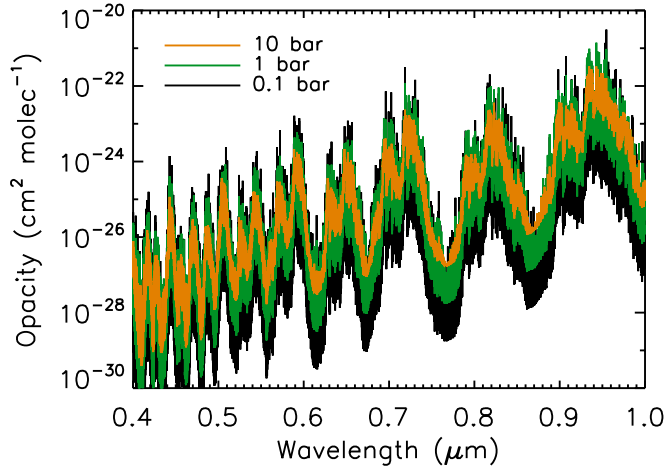
Figure 3. Left: high-resolution ($1\ cm^{-1}$) H_2O opacities from 0.4 to 1.0 μm at three different pressures: 0.1, 1, and 10 bars. Right: absorption features in an $R = 140$ spectrum from 0.3 to 1.05 μm of H_2O, O_2, and O_3 at the fiducial mixing ratios listed in Table 1 at $P = 1$ and 10 bars. For each spectrum here, the atmosphere only contains the stated molecule and a radiatively inactive filler gas to match the pressure.

of the output spectra. For each model layer, we interpolate over the opacities from our table given the pressure. The chemical abundances in our forward-model atmosphere are constant as a function of pressure, and we also adopt a uniform acceleration due to gravity.

We have also added an option to include partial cloudiness across a planetary disk whose fractional coverage is described by f_c. To mimic partial cloudiness as we see on Earth, we call the forward model twice. We use the same set of atmospheric and planetary parameters for both calls, except for the cloud optical depth. "Cloudy" is the call that has a nonzero cloud optical depth, while "cloud-free" is the call where we set the cloud optical depth to zero. Each call returns a geometric albedo spectrum, and we combine the two sets with the fractional cloudiness parameter such that the combined spectrum follows $f_c \times$ cloudy $+ (1 - f_c) \times$ cloud–free.

2.3. Albedo Model Fiducial Values and Validation

The generalized three-dimensional albedo model described above can simulate reflected-light spectra of a large diversity of planet types, spanning solid-surfaced worlds to gas giants with a variety of prescribed atmospheric compositions. For the present study, however, we choose to focus on Earth-like worlds, which are described in detail below. Thus, we define a set of fiducial model input parameters that are designed to mimic Earth and thereby enable us to generate simulated observational data sets for an Earth twin.

Table 1 summarizes the fiducial model parameter values adopted for our Earth twin. Also shown are the priors for these parameters, which we use when performing retrieval analyses. For an Earth-like setup, the surface atmospheric pressure is $P_0 = 1$ bar, and we adopt a surface albedo of $A_s = 0.05$, which is representative of mostly ocean-covered surface. We adopt a uniform acceleration due to gravity of $g = 9.8\ m\ s^{-2}$ and set the planetary radius to $R_\oplus$. For convenience, we sometime refer to these four variables (P_0, A_s, g, and R_p) as the bulk planetary and atmospheric parameters.

We focus on molecular absorption due to H_2O, O_3, and O_2. While our albedo model includes opacities from CH_4 and CO_2 as well, we omit these two species, as the reflected-light spectrum of Earth in the visible contains no strong features for these molecules. The input values for the molecular

Table 1
List of the 11 Retrieved Parameters in the Complete Cloudy Model, Their Descriptions, Fiducial Input Values, and Corresponding Priors

Parameter	Description	Input	Prior
$\log P_0$ (bar)	Surface pressure	$\log(1)$	$[-2,2]$
$\log H_2O$	Water vapor mixing ratio	$\log(3 \times 10^{-3})$	$[-8,-1]$
$\log O_3$	Ozone mixing ratio	$\log(7 \times 10^{-7})$	$[-10,-1]$
$\log O_2$	Molecular oxygen mixing ratio	$\log(0.21)$	$[-10,0]$
R_p ($R_\oplus$)	Planet radius	1	$[0.5, 12]$
$\log g$ (m s^{-2})	Surface gravity	$\log(9.8)$	$[0,2]$
$\log A_s$	Surface albedo	$\log(0.05)$	$[-2, 0]$
$\log p_t$ (bar)	Cloud-top pressure	$\log(0.6)$	$[-2,2]$
$\log dp$ (bar)	Cloud thickness	$\log(0.1)$	$[-3,2]$
$\log \tau$	Cloud optical depth	$\log(10)$	$[-2,2]$
$\log f_c$	Cloudiness fraction	$\log(0.5)$	$[-3,0]$

abundances (or volume mixing ratios) are $H_2O = 3 \times 10^{-3}$, $O_3 = 7 \times 10^{-7}$, and $O_2 = 0.21$. These abundance values are based on column-weighted averages from a standard Earth model atmosphere with vertically varying gas mixing ratios (McClatchey et al. 1972). The primary Rayleigh scatterer and background gas in our fiducial model is N_2, whose abundance makes up the remainder of the atmosphere after all other gases are accounted for (i.e., roughly 0.79). Rayleigh scattering is treated according to Hansen & Travis (1974) with constants to describe the scattering properties of N_2, O_2, and H_2O from Allen & Cox (2000). We do not include polarization or Raman-scattering effects.

Our cloud model was designed to be minimally parametric while still enabling us to sufficiently reproduce realistic spectra of Earth. Our single-layer gray H_2O cloud has $\bar{\omega} = 1$ and $\bar{g} = 0.85$, which are characteristic of water clouds across the visible range. These two parameters were fixed to minimize retrieval model complexity, as we believe that water is the most likely condensate for worlds in the habitable zone. Nevertheless, future studies may not wish to assume values of $\bar{\omega}$ and $\bar{g}$ a priori. Cloud-top pressure (p_t) and fractional coverage (f_c) are set at 0.6 bar and 50%, respectively, which are roughly consistent with observations of optically thick cloud coverage on Earth (Stubenrauch et al. 2013). Cloud thickness (dp) and
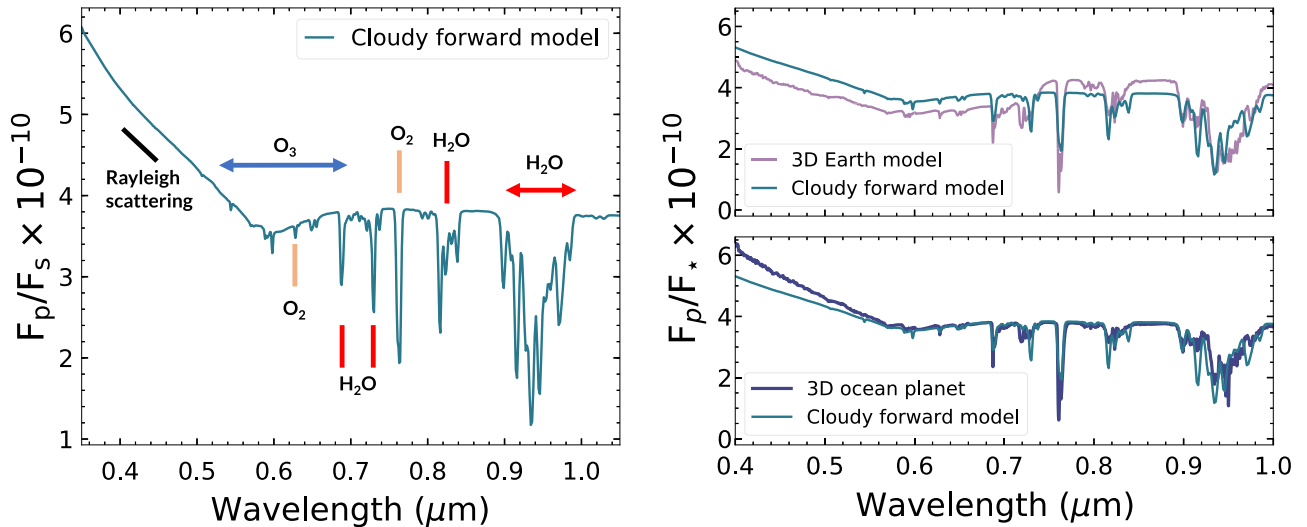
Figure 4. Left: spectrum generated with the forward model in this study using fiducial values from Table 1. Key spectral features from the atmospheric species in our model are labeled. Right, top: comparison of the cloudy forward model in this study using fiducial values from Table 1 to a spectrum from a more computationally complex 3D forward model of Earth at full phase described in Robinson et al. (2011). Right, bottom: comparison of the cloudy forward model to a spectrum of a planet generated using the 3D model from Robinson et al. (2011) that is like Earth except it only has ocean coverage.

optical depth (τ) were set to 0.1 bar and 10, respectively, based on results from the MODIS instrument (http://modis-atmos.gsfc.nasa.gov) used in Robinson et al. (2011).

With fiducial values chosen, we validate our forward model against a simulated high-resolution disk-integrated spectrum of Earth at full phase, as shown in Figure 4. The comparison spectrum is produced by the NASA Astrobiology Institute's Virtual Planetary Laboratory (VPL) sophisticated 3D, line-by-line, multiple-scattering spectral Earth model (Robinson et al. 2011). The Robinson et al. (2011) tool can simulate images and disk-integrated spectra of Earth from the ultraviolet to the infrared. It has been validated against observations at visible wavelengths taken by NASA's *EPOXI* (Robinson et al. 2011) and *LCROSS* missions (Robinson et al. 2014).

Features of the Robinson et al. (2011) model include Rayleigh scattering due to air molecules, realistic patchy clouds, and gas absorption from a variety of molecules, including H_2O, CO_2, O_2, O_3, and CH_4. Surface coverage of different land types (e.g., forest, desert) is informed by satellite data, and water surfaces incorporate specular reflectance of sunlight. A grid of thousands of surface pixels is nested beneath a grid of 48 independent atmospheric pixels, all of equal area. For each surface pixel, properties from the overlying atmospheric pixels are used as inputs to a full-physics, plane-parallel radiative transfer solver: the Spectral Mapping Atmospheric Radiative Transfer (SMART) model (Meadows & Crisp 1996). Intensities from this solver are integrated over the pixels with respect to solid angle, thereby returning a disk-integrated spectrum.

The sophistication of the Robinson et al. (2011) model makes it unsuitable to retrieval studies, however, as model runtimes are measured in weeks for the highest-complexity scenarios. This, in part, justifies our adoption of a minimally parametric albedo model (whose runtime is measured in seconds). Furthermore, as in Figure 4, our efficient albedo model reproduces all of the key features of the Robinson et al. (2011) model. The most notable differences—that the efficient model, as compared to the Robinson et al. (2011) model, is more reflective in the blue and less reflective in the red—are simply due to our adoption of a gray surface albedo. Land and

plants, which cover roughly 29% of Earth's surface, are generally more reflective in the red than in the blue. Figure 4 also compares a spectrum from our forward model against a spectrum of a partially clouded ocean planet generated with the Robinson et al. (2011) model. This ocean world is identical to Earth except for the fact that its surface is covered entirely by an ocean, with no land present. The surface albedo in the ocean model is gray beyond 500 nm; shortward of this, the reflectivity increases, likely leading to the discrepancy in our comparison at the bluest wavelengths. Still, with a more accurate match to a planet that has a nearly gray albedo through the visible, we consider our assumption of gray surface albedo to be the main reason for the discrepancies when compared to the Robinson et al. (2011) realistic model.

Finally, in our albedo model, we set 100 facets for the visible hemisphere and calculate a high-resolution geometric albedo spectrum at 1000 wavelength points between 0.35 and 1.05 μm. Like Lupu et al. (2016), we only consider a planet at full phase ($\alpha = 0°$). While direct-imaging missions will not obtain observations of exoplanets at full phase, this assumption makes little difference for our results, as we are not computing integration times and only work in S/N space. Additionally, as Nayak et al. (2017) followed up Lupu et al. (2016) by retrieving phase information from giant planets in reflected light, we anticipate performing a similar expansion in the future. Our forward model has 61 pressure levels in an isothermal atmosphere of 250 K, bounded below by a reflective surface. The top of the atmosphere is set at $P_{\rm top} = 10^{-4}$ bars.

2.4. Retrieval Setup and Noise Model

We convert a high-resolution geometric albedo spectrum to a synthetic planet-to-star flux ratio spectrum given the resolution of an instrument and a noise model. We then apply a Bayesian inference tool on the synthetic data set to sample the posterior probability distributions of the forward-model input parameters. To perform Bayesian parameter estimation, we utilize the open-source affine-invariant MCMC ensemble sampler `emcee` (Goodman & Weare 2010; Foreman-Mackey et al. 2013). Ensemble refers to the use of many chains, or walkers,

to traverse parameter space; as a massively parallelized algorithm, it is computationally efficient. Affine invariance refers to the invariant performance under linear transformations of parameter space, enabling the algorithm to be insensitive to parameter covariances (Foreman-Mackey et al. 2013). With a cloudy retrieval, we can expect complex correlations that a sampler should be able to reveal. As it is more agnostic to the shape of the posterior, we choose `emcee` following Nayak et al. (2017) over `Multinest`, another sampler Lupu et al. (2016) considered that yielded consistent results. The albedo model is coded in Fortran; we convert it into a Python-callable library with the F2PY package. Each call to the forward model takes approximately 10 s of clock time on an eight-core processor. To visualize the MCMC results, we utilize the `corner`-plotting package developed by Foreman-Mackey (2016).

Table 1 lists the priors for our parameters. We offer a generous range on the molecular abundances; we allow O_2 in particular to dominate the atmospheric composition. Our choice of radius range (0.5–12 $R_\oplus$) reflects the range of planetary sizes from Mars to Jupiter. Also, when performing retrievals, we impose two limiting conditions to maintain physical scenarios. First, we limit the mixing ratio of N_2, $f_{N_2} = 1 - \sum$ (gas abundances), to be between 0 and 1. Second, for the cloud pressure terms, we reject any drawn value that does not satisfy $10^{P_t} + 10^{dp} < 10^{P_0}$ (i.e., that the cloud base cannot extend below the bottom of the atmosphere). Note that for the purposes of the retrieval, we consider pressures in log space.

We simulate noise in our observations following the expressions given in Robinson et al. (2016). For simplicity, we include only read noise and dark current, as Robinson et al. (2016) showed that detector noise will be the dominant noise source in *WFIRST*-type spectral observations of exoplanets. Detector and instrument parameters for the *HabEx* and *LUVOIR* concepts are only loosely defined, and advances in detector technologies for these missions may move observations out of the detector-noise-dominated regime. In the detector-noise-dominated regime, the S/N is simply

$$\text{S/N} = \frac{c_p \times t_{\text{int}}}{\sqrt{(c_d + c_r) \times t_{\text{int}}}} \qquad (13)$$

,where t_{int} is the integration time, c_p is the planet count rate, c_d is the dark-noise count rate, and c_r is the read-noise count rate. More rigorously, it can be shown that, at constant spectral resolution, $\text{S/N} \propto q \mathcal{T} A_g \Phi(\alpha) B_\lambda \lambda$, where q is the wavelength-dependent detector quantum efficiency, $\mathcal{T}$ is throughput, and B_λ is the host stellar specific intensity (taken here as a Planck function at the stellar effective temperature). We use a stellar temperature of 5780 K for the blackbody. When the S/N at one wavelength is specified, this scaling implies that the calculation of the S/N at other wavelengths is independent of the imaging raw contrast of the instrument. We can expect the noise at the redder end of our range to be large, as the detector quantum efficiency (taken to be appropriate for the *WFIRST*/CGI) rapidly decreases. Since we treat only S/N rather than modeling exposure times, the exact mix of noise sources is not relevant (so that, e.g., dark current and read noise are indistinguishable). The key relevant properties of the noise model are that it is uncorrelated between spectral channels and

its magnitude only depends on wavelength via a dependence on point-spread function area (Robinson et al. 2016, their Equation (26)), which will be true for detector-limited cases but may not be true for large-aperture instruments limited by speckle noise.

For our study, we will consider multiple wavelength resolutions, *R*, and S/Ns. Working in S/Ns (instead of integration times) makes our investigations independent of telescope diameter, target distance, and other system-specific or observing parameters. Because the S/N is dependent on wavelength, we reference our values to be at the *V* band (550 nm) for all resolutions for *HabEx*/*LUVOIR*. Since the *WFIRST*/CGI spectrograph is currently planned to only extend to 600 nm at the blue end, we opt to reference our *WFIRST* S/Ns to this wavelength. Unlike previous studies (Lupu et al. 2016; Nayak et al. 2017), our simulated *WFIRST* rendezvous data include two photometric points in the blue, which is consistent with current CGI designs. We set the S/N in the *WFIRST* filters to be equal to that at 600 nm.

Our simulation grid setup is shown in Table 2, where the spectral resolutions and S/Ns assumed for different observing scenarios are indicated. Figure 5 demonstrates the *WFIRST* rendezvous scenario data along with $R = 70$ and 140 data points (for *HabEx*/*LUVOIR*) plotted over the forward-model spectrum before noise is added. The scaling of the S/N with wavelength for the *WFIRST* rendezvous (normalized to unity at 600 nm), as well as our $R = 70$ and 140 cases (normalized to unity at 550 nm), is shown in Figure 6. The impact of the host stellar SED sets the overall shape of the S/N scaling, with additional influence from atmospheric absorption band features, as well as detector quantum efficiency effects (that have strong impacts at red wavelengths). Thus, Figure 6 can be used to translate our stated S/N to the S/N at any other wavelength (e.g., an S/N = 10 simulation has an S/N in the continuum shortward of the 950 nm water vapor band of roughly $0.3 \times 10 = 3$).

When generating simulated data with a noise model, there are several options for handling the placement/sampling of the mock observational data points. Previous studies (Lupu et al. 2016; Nayak et al. 2017) have generated a single randomized data set for a given S/N. The placement of a single spectral data point is determined by randomly sampling a Gaussian distribution whose width is determined by the wavelength-dependent S/N. While this treatment can accurately simulate a single observational instance, it also runs the risk (especially at lower spectral resolution and S/N) of biasing retrieval results, as the random placement of only a small handful of spectral data points can significantly impact the outcome. Given this, it is ideal to retrieve on a large number ($\gtrsim 10$) of simulated data sets at a given spectral resolution and S/N, where a comprehensive view of all the posteriors from the collection of instances will indicate expected telescope/instrument performance. Unfortunately, given the large number of *R*–S/N pairs in our study (10) and the long runtime of an individual retrieval (of order 1 week on a cluster), running ~10 noise instances for each of our *R*–S/N pairs is computationally unfeasible (requiring ~100 weeks of cluster time). Thus, we opt for an intermediate approach that maintains computational feasibility and avoids potential biases from individual noise instances. Here we run only a single noise instance at a given *R*–S/N pair, but we do not randomize the placement of the individual spectral points. In other words, the individual
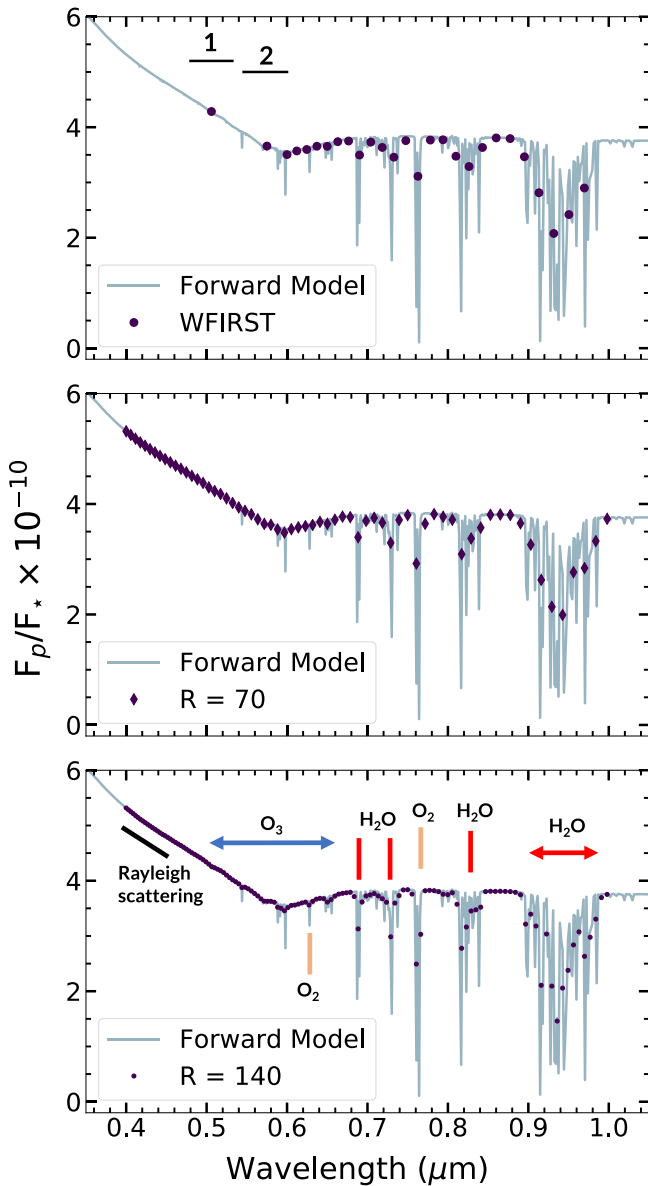
Figure 5. High-resolution (1000 wavelength points from 0.35 to 1.05 μm) forward-model spectrum, overplotted with simulated *WFIRST* rendezvous and $R = 70$ and 140 data (top to bottom). Key spectral features for atmospheric gases in our model are labeled. In the top panel, "1" and "2" mark the span of the *WFIRST* Design Cycle 7 filters (see Table 2).

Table 2
Simulated Data Sets

	$R = 70, R = 140$	*WFIRST* Rendezvous[a]
Wavelength (μm)	0.4–1.0	0.506, 0.575[b], $R = 50$: 0.6–0.97[c]
Data quality	$S/N_{550nm} = 5, 10, 15, 20$	$S/N_{600nm} = 5, 10, 15, 20$

Note. We do not randomize the noise for any of the data sets.
[a] Using *WFIRST* Design Cycle 7 values from https://wfirst.ipac.caltech.edu/sims/Param_db.html.
[b] The first photometric band is centered on 0.506 μm and covers 0.48–0.532 μm. The second photometric band is centered on 0.575 μm and covers 0.546–0.6 μm. We assume 100% transmission.
[c] We combine three integral field spectrograph bands into one at $R = 50$ from 0.6 to 0.97 μm. Separated, they are 0.6–0.72, 0.7–0.84, and 0.81–0.97 μm.

3. Retrieval Validation

Before using our framework on simulated data, we validate its accuracy and examine its performance. For this initial validation, we use nonrandomized, wavelength-independent noise at an S/N of 20 for a spectrum at a resolution of $R = 140$. Table 3 lists our four validation model variants, each increasing in complexity as we systematically explore how the addition of retrieved parameters influences the posterior distributions and correlations. In Model I, we fix all parameters except P_0 and A_s. In Model II, we add g and R_p; in Model III, we add gases as retrieved parameters (H$_2$O, O$_3$, O$_2$); and in Model IV, we add all cloud parameters. Incrementally increasing the number of free parameters (from 2 to 11) allows us to see the interconnections between them and helps us understand how clouds can obscure our inferences.

In Figure 7, we present the posterior distributions for Model I. In the two-dimensional correlation histogram, a higher probability corresponds to a darker shade. With all else held constant, we see narrow posterior distributions and a slight correlation between P_0 and A_s. For lower values of surface pressure, which controls the turnoff of the Rayleigh-scattering slope, we need a brighter surface to maintain the measured brightness, especially in the red end of the spectrum, and vice versa. We mark the 16%, 50%, and 84% quantiles in the marginalized one-dimensional posterior distributions. The posterior distributions for Model II are shown in Figure 8 and are generally narrow (as only four parameters are being retrieved). There are two key correlations, one between g and P_0 and one between R_p and A_s. Both gravity and surface pressure influence the column mass so that, when attempting to fit a spectrum, we can trade a larger gravity with a larger surface pressure and maintain a similar column mass (which controls, e.g., the Rayleigh-scattering feature). Additionally, we can trade off a larger reflecting surface area (i.e., larger R_p) with a darker surface (lower A_s), which is a statement of the typical "radius/albedo degeneracy" problem. The posterior for surface albedo is now an upper limit instead of a constraint. As a result, the radius posterior distribution appears truncated at larger values given the tight correlation between these two parameters. The correlation seen originally in Model I, between P_0 and A_s, then acts as a chain between the other two more prominent correlations to induce correlations between parameters such as A_s and g or R_p and P_0.

Once we allow gases to be free parameters in Model III (Figure 9), the P_0 and A_s correlation becomes diminished as

simulated spectral points are placed on the "true" planet-to-star flux ratio point and assigned error bars according to the S/N and noise model. While this approach prevents having a small handful of randomized data points biasing the retrieval results, it does lead to likely optimistic results, especially at modest S/N (i.e., S/N ~ 10), since data point randomization is, in effect, an additional "noise" source that we are omitting. This means that the posterior distributions will usually be centered on the true values in an unrealistic fashion. However, the width and shape of the posterior covariances will be representative of real observations, so the fidelity of retrievals can be assessed. We keep this optimism in mind when discussing results in later sections; in particular, we compare the performance of retrievals on multiple noise instances of a subset of the cases we consider to the nonrandomized case in Section 5.2.
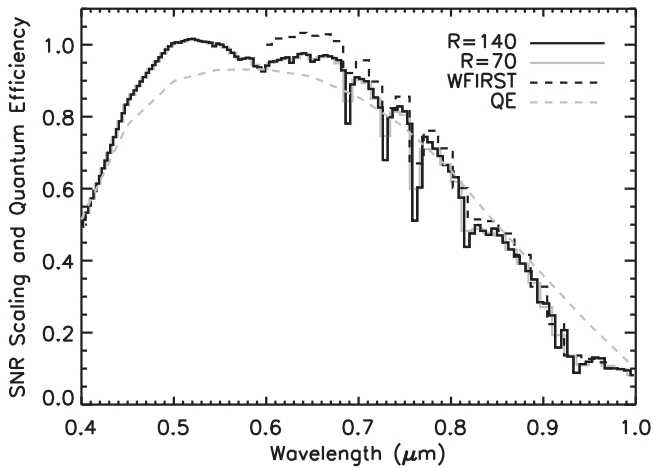
Figure 6. Scaling of S/N with wavelength for *WFIRST* rendezvous, $R = 70$, and $R = 140$ cases. The *WFIRST* curve is normalized to unity at 600 nm, while the $R = 70$ and 140 curves are normalized to unity at 550 nm, following our definition of simulation S/N at these respective wavelengths. Also shown is the wavelength-dependent detector quantum efficiency (QE) that we adopt.

Table 3
Four Cumulative Models for Retrieval Validation, as Described in Section 3

Model Variant		Retrieved Parameters	N_{param}
I	Surface conditions	P_0, A_s	2
II	+ Bulk properties	P_0, A_s, g, R_p	4
III	+ Gas mixing ratios	P_0, A_s, g, R_p	7
		H_2O, O_2, O_3	
IV	+ Cloud properties	P_0, A_s, g, R_p	11
		H_2O, O_2, O_3	
		p_t, dp, τ, f_c	

Notes. See Table 1 for the corresponding definition and prior of each parameter. Model IV represents the full suite of parameters and can serve as a reference for the fixed parameters in Models I through III.

H_2O, due to its numerous bands across the spectral range, becomes a primary control of brightness. The significant impact of H_2O on the spectrum leads to a strong positive correlation between H_2O abundance and planetary size, as additional water vapor absorption can be compensated by a larger planetary size to maintain fixed brightness. We now see gravity linked to the molecular abundances, which is expected, as surface gravity directly influences the column abundance of a species. This key correlation also causes the individual gas abundances to be correlated with each other. The main correlations from Model II are still present. We note once more that we do not have constraints on the surface albedo, again leading to an asymmetric distribution for radius. Thus, from the strong correlation of H_2O with R_p and the fundamental correlation between R_p and A_s, we see correlations between planetary radius, surface albedo, and all gas abundances. Weak correlations between surface pressure and gas abundances are due to column abundance effects.

Finally, as shown by Figure 10, we retrieve on the data with the full forward model, adding in the cloud parameters p_t, dp, τ, and f_c. This version of the model is what we apply when simulating direct-imaging data in the upcoming sections and represents our most realistic (i.e., true to the actual Earth) scenario. The optical depth is shown to only have a lower-limit
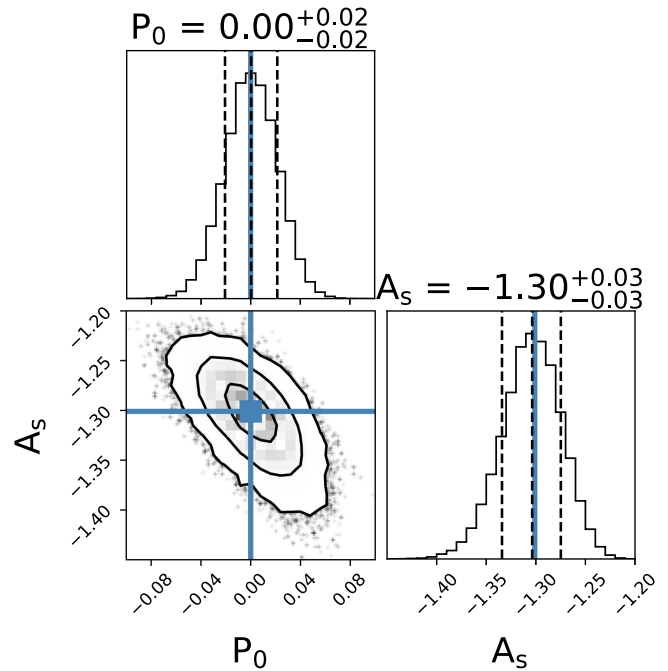

Figure 7. Posterior distributions of Model I from Table 3, where we fix all parameters but P_0 and A_s. We retrieve on $R = 140$, S/N = 20 data with wavelength-independent noise. Overplotted in light blue are the fiducial parameter values. The 2D marginalized posterior distribution, used in interpreting correlations, is overplotted with the 1σ, 2σ, and 3σ contours. Above the 1D marginalized posterior for each parameter, we list the median retrieved value with uncertainties that indicate the 68% confidence interval. Dashed lines (left to right) mark the 16%, 50%, and 84% quantiles.

constraint. Thus, the retrieval detects a cloud but cannot constrain the optical depth beyond showing that the cloud is optically thick. There is an expected correlation between τ and f_c; a higher cloudiness fraction can complement a less optically thick cloud, and vice versa. There is only an upper limit to dp, which is a result of the lack of vertical sensitivity given the constant-with-pressure abundance distributions. The posterior distribution for O_2 becomes a lower limit instead of a constraint, as in Model III. Surface gravity is less precisely and accurately constrained compared to the previous, less complex renditions of the model.

For optically thin clouds, we expect to better constrain surface albedo; however, we do not consider this scenario in our study. We examined instead the performance of a completely cloud-free model on data generated with our cloudy model. We find that while the model can fit the data and return accurate estimates of, e.g., the mixing ratios, we get inaccurate estimates of the surface albedo and pressure. These two parameters are biased, with lower surface pressure paired with higher surface albedo as the preferred configuration in the cloud-free case. As a result, we move forward with utilizing our cloudy forward model on our simulated data. However, we note that in realistic cases where we do not know the true state of a planet's atmosphere, we could obtain complementary information relating to the presence of clouds (e.g., variability) such that we may choose the most appropriate forward model.

4. Results

We generate data sets for *HabEx*- and *LUVOIR*-like missions (0.4–1.0 μm at $R = 70$ and 140) at S/N = 5, 10, 15, and 20 and for the *WFIRST* rendezvous scenario (two photometric
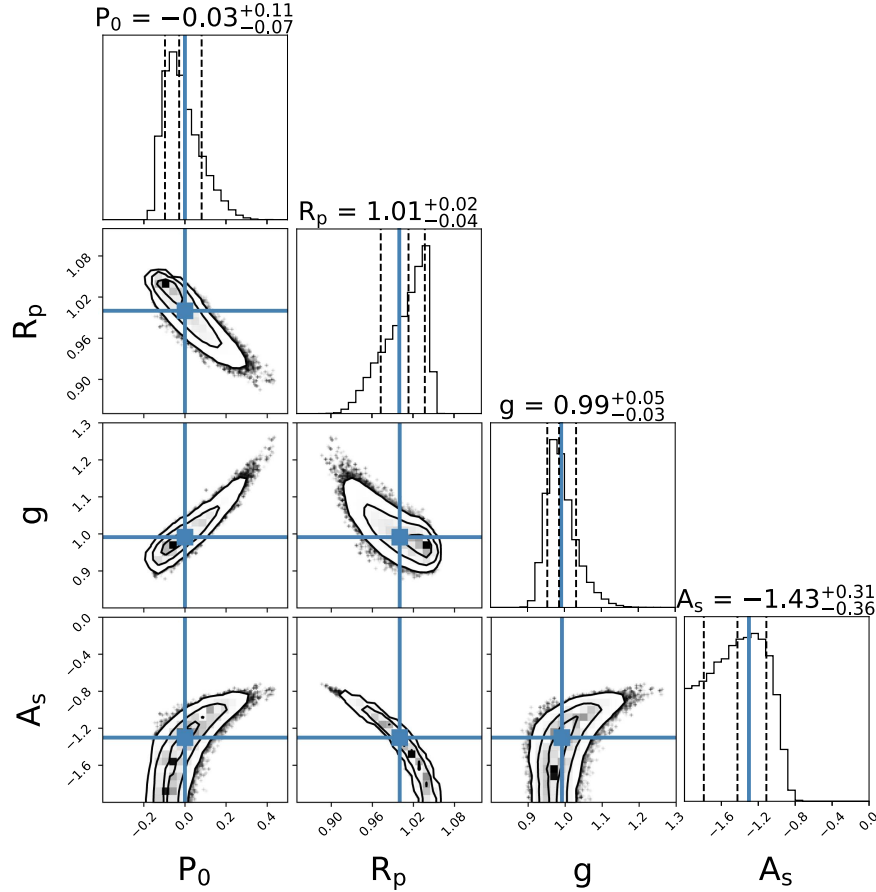
Figure 8. Posterior distributions of Model II from Table 3, where we fix all parameters except for P_0, A_s, g, and R_p. We retrieve on $R = 140$, $S/N = 20$ data with wavelength-independent noise. Overplotted in light blue are the fiducial parameter values. The 2D marginalized posterior distribution, used in interpreting correlations, is overplotted with the 1σ, 2σ, and 3σ contours. Above the 1D marginalized posterior for each parameter, we list the median retrieved value with uncertainties that indicate the 68% confidence interval. Dashed lines (left to right) mark the 16%, 50%, and 84% quantiles.

points within 0.4–0.6 μm plus a spectrum of $R = 50$ for 0.6–0.96 μm), also at $S/N = 5$, 10, 15, and 20. In all cases, we use the noise model to generate the uncertainties expected for high-contrast imaging instead of the wavelength-independent noise for the validations in the previous section. As Section 2.4 describes, the S/N refers to the value at 0.55 μm for $R = 70$ and 140 and at 0.6 μm for *WFIRST*. We record the specific runs in Table 2. In place of showing the correlations for all parameters for all cases, we refer to Figure 10, which represents the ideal case correlations among the parameter posteriors. We only show the posterior probability distributions themselves to better highlight any trends with respect to S/N and/or R. We grouped the posteriors in terms of, first, bulk atmospheric and planetary parameters (P_0, R_p, g, A_s), then cloud parameters (p_t, dp, τ, f_c), and finally gases (H_2O, O_3, O_2). For each case, `emcee` was run with 16 MCMC chains (walkers) per parameter for at least 12,000 steps, the last 5000 of which are used to determine the posterior distributions. From those 5000 steps, we randomly selected 1000 sets of parameters to calculate their corresponding high-resolution spectra. These spectra are plotted with the data to show the 1σ, 2σ, and median fits.

4.1. Results for R = 70 and 140 Simulated Data

For both $R = 70$ and $R = 140$, we simulated data sets at $S/N = 5$, 10, 15, and 20. Table 4 lists the median and 1σ

values of all retrieved parameters for each S/N at $R = 70$. Figure 11 shows the marginalized posterior distributions for the model parameters for all S/N cases for $R = 70$, plotted with the fiducial or "truth" values. Table 5 lists the median and 1σ values of all retrieved parameters for each S/N at $R = 140$. Figure 12 shows the posterior distributions for $R = 140$ for the model parameters for all S/N cases compared against their input values. Figure 14 shows the corresponding spread in fits and the median fit to the data for each S/N for both resolutions.

4.2. Results for WFIRST Rendezvous Simulated Data

For the *WFIRST* rendezvous scenario, we utilized the Design Cycle 7 instrument parameters to set the locations of the two photometric points and the range and resolution of the spectrometer ($R = 50$; see Table 2). Because of this particular setup, we reference the S/Ns in our grid (5, 10, 15, 20) at 600 nm and assign the photometric points the same S/N as at 600 nm. Table 6 lists the median and 1σ values of all retrieved parameters for each S/N variant. Figure 13 presents the posterior distributions for the four *WFIRST* rendezvous variants with respect to the input values. Figure 14 shows the spread in fits and median fit to the data for each variant.

5. Discussion

The results from our retrieval analyses enable us to identify the S/N required at a given spectral resolution to constrain key
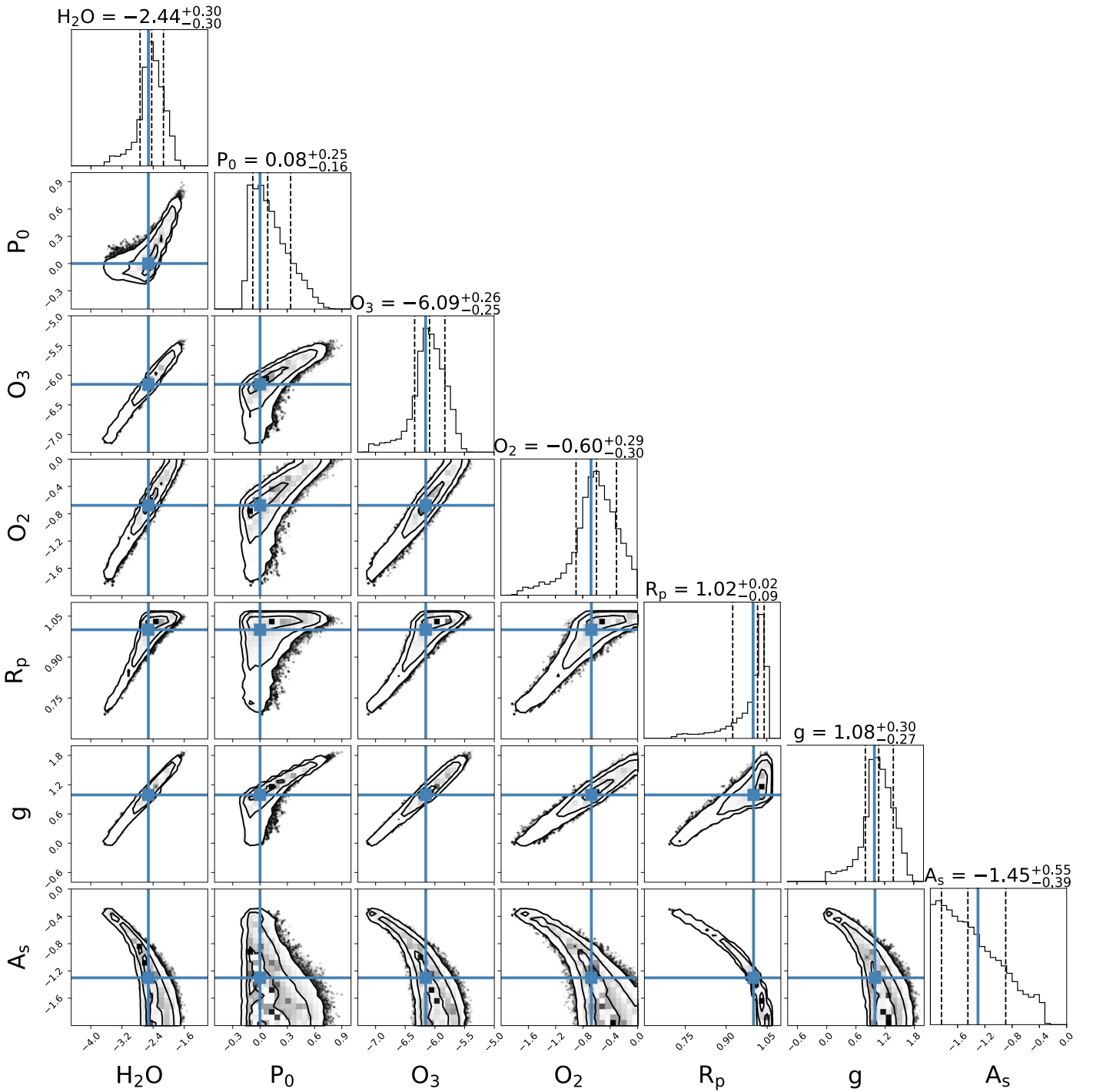
Figure 9. Posterior distributions of Model III from Table 3, where we retrieve P_0, A_s, g, R_p, H_2O, O_2, and O_3. We retrieve on $R = 140$, S/N $= 20$ data with wavelength-independent noise. Overplotted in light blue are the fiducial parameter values. The 2D marginalized posterior distribution, used in interpreting correlations, is overplotted with the 1σ, 2σ, and 3σ contours. Above the 1D marginalized posterior for each parameter, we list the median retrieved value with uncertainties that indicate the 68% confidence interval. Dashed lines (left to right) mark the 16%, 50%, and 84% quantiles.

planetary and atmospheric quantities. These findings have important implications for the development of future space-based direct-imaging missions. We discuss these ideas below and also touch on impacts of certain model assumptions and ideas for future research directions.

In what follows, we define a "weak detection" for a given parameter as having a posterior distribution that has a marked peak but that also has a substantial tail toward extreme values

(indicating that, e.g., for a gas, we could not definitively state that the gas is present in the atmosphere). A "detection" implies a peaked posterior without tails toward extreme values but whose 1σ width is larger than an order of magnitude. We use the term "constraint" to indicate a detection whose posterior distribution has a 1σ width smaller than an order of magnitude. A nondetection would be indicated by a flat posterior distribution across the entire (or nearly entire) prior range.
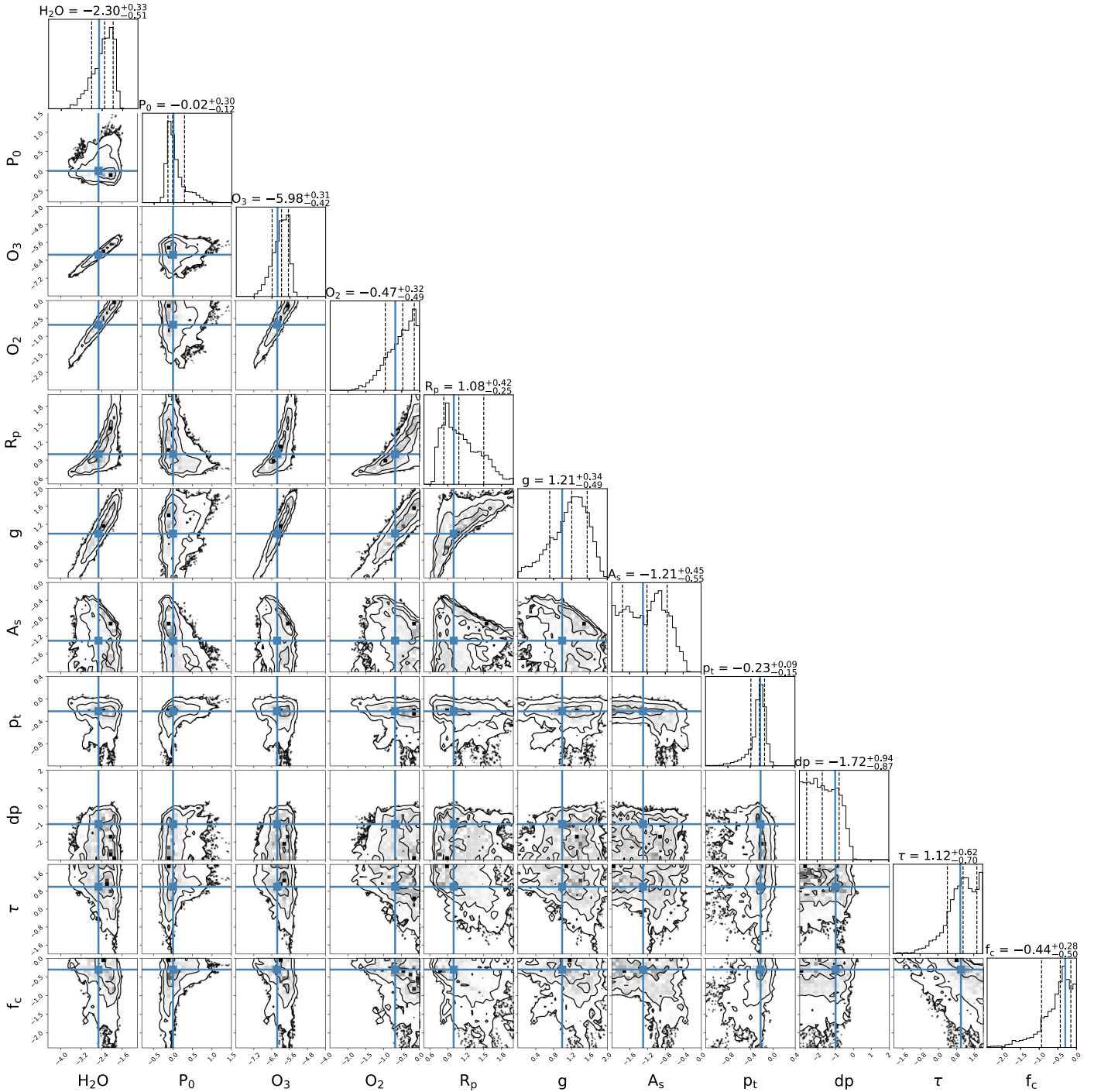
Figure 10. Posterior distributions of Model IV, or the complete model, from Table 3. We retrieve for 11 parameters: P_0, A_s, g, R_p, H_2O, O_2, O_3, p_t, dp, τ, and f_c. We retrieve on $R = 140$, S/N = 20 data with wavelength-independent noise. Overplotted in light blue are the fiducial parameter values. The 2D marginalized posterior distribution, used in interpreting correlations, is overplotted with the 1σ, 2σ, and 3σ contours. Above the 1D marginalized posterior for each parameter, we list the median retrieved value with uncertainties that indicate the 68% confidence interval. Dashed lines (left to right) mark the 16%, 50%, and 84% quantiles.

For planetary radius, which is not retrieved in logarithmic space, we distinguish between a "detection" and a "constraint" when the 1σ uncertainties are small enough to firmly place the planet in the Earth/super-Earth regime (i.e., with a radius below 1.5 $R_\oplus$; Rogers 2015; Chen & Kipping 2017). A visual summary of weak detections, detections, and constraints as a function of S/N for our different observing scenarios and a selection of key parameters is given in Tables 7–9.

5.1. Influence of S/N on Inferred Properties

For $R = 70$ at S/N = 5, Figure 11 shows that there is only a weak detection of P_0 and a detection of R_p, which merely suggests that the planet has an atmosphere and is not a giant planet. As S/N increases to 10, the O_3 posterior distribution has a weak peak near the fiducial value, and the gas is only weakly detected. Once the S/N is equal to 15, we weakly detect H_2O, O_3, and O_2. At an S/N of 20, it is possible to detect

13

Table 4
$R = 70$ Retrieval Results, with Median Value and 1σ Uncertainties of the Parameters

Parameter	Input	S/N = 5	S/N = 10	S/N = 15	S/N = 20
log H_2O	−2.52	$-5.07^{+2.34}_{-1.92}$	$-3.85^{+1.77}_{-2.60}$	$-3.12^{+0.97}_{-1.71}$	$-2.76^{+0.62}_{-0.88}$
log O_3	−6.15	$-7.55^{+1.49}_{-1.46}$	$-6.79^{+0.93}_{-1.81}$	$-6.37^{+0.55}_{-0.84}$	$-6.24^{+0.47}_{-0.60}$
log O_2	−0.68	$-5.12^{+3.25}_{-3.23}$	$-4.51^{+3.24}_{-3.61}$	$-1.86^{+1.29}_{-3.99}$	$-1.00^{+0.66}_{-1.01}$
log P_0	0.0	$0.02^{+1.35}_{-0.84}$	$-0.03^{+0.87}_{-0.70}$	$0.28^{+0.85}_{-0.56}$	$0.25^{+0.56}_{-0.49}$
R_p	1.0	$1.23^{+1.54}_{-0.58}$	$1.33^{+1.52}_{-0.52}$	$0.97^{+0.68}_{-0.27}$	$0.98^{+0.44}_{-0.25}$
log g	0.99	$1.33^{+0.48}_{-0.77}$	$1.48^{+0.38}_{-0.68}$	$1.28^{+0.51}_{-0.66}$	$1.24^{+0.55}_{-0.69}$
log A_s	−1.3	$-0.96^{+0.58}_{-0.74}$	$-1.05^{+0.55}_{-0.59}$	$-0.70^{+0.37}_{-0.62}$	$-0.63^{+0.29}_{-0.46}$
log p_t	−0.22	$-1.14^{+0.97}_{-0.61}$	$-1.19^{+0.93}_{-0.56}$	$-0.92^{+0.86}_{-0.71}$	$-0.94^{+0.84}_{-0.73}$
log dp	−1.0	$-1.67^{+1.24}_{-0.92}$	$-1.71^{+1.18}_{-0.91}$	$-1.35^{+1.17}_{-1.14}$	$-1.43^{+1.11}_{-1.06}$
log τ	1.0	$0.10^{+1.30}_{-1.43}$	$0.21^{+1.23}_{-1.48}$	$0.49^{+1.03}_{-1.66}$	$0.61^{+0.93}_{-1.66}$
log f_c	−0.3	$-1.43^{+0.99}_{-1.07}$	$-1.33^{+0.94}_{-1.12}$	$-0.93^{+0.71}_{-1.32}$	$-1.05^{+0.80}_{-1.27}$

Table 5
$R = 140$ Retrieval Results, with Median Value and 1σ Uncertainties of the Parameters

Parameter	Input	S/N = 5	S/N = 10	S/N = 15	S/N = 20
log H_2O	−2.52	$-4.56^{+2.14}_{-2.35}$	$-2.74^{+0.69}_{-1.07}$	$-2.61^{+0.47}_{-1.07}$	$-2.43^{+0.39}_{-0.56}$
log O_3	−6.15	$-7.36^{+1.26}_{-1.65}$	$-6.26^{+0.53}_{-0.68}$	$-6.18^{+0.42}_{-0.48}$	$-6.03^{+0.34}_{-0.48}$
log O_2	−0.68	$-4.45^{+3.08}_{-3.69}$	$-1.06^{+0.76}_{-1.43}$	$-0.76^{+0.51}_{-0.79}$	$-0.60^{+0.43}_{-0.59}$
log P_0	0.0	$0.07^{+1.01}_{-0.84}$	$0.20^{+0.72}_{-0.49}$	$0.12^{+0.49}_{-0.36}$	$0.07^{+0.39}_{-0.31}$
R_p	1.0	$1.25^{+1.16}_{-0.52}$	$1.01^{+0.60}_{-0.28}$	$0.99^{+0.42}_{-0.23}$	$1.05^{+0.42}_{-0.27}$
log g	0.99	$1.36^{+0.46}_{-0.74}$	$1.31^{+0.49}_{-0.77}$	$1.14^{+0.56}_{-0.65}$	$1.20^{+0.50}_{-0.64}$
log A_s	−1.3	$-0.98^{+0.54}_{-0.60}$	$-0.67^{+0.32}_{-0.50}$	$-0.68^{+0.29}_{-0.44}$	$-0.79^{+0.34}_{-0.69}$
log p_t	−0.22	$-1.23^{+1.03}_{-0.55}$	$-0.96^{+0.80}_{-0.71}$	$-0.79^{+0.70}_{-0.82}$	$-0.66^{+0.53}_{-0.85}$
log dp	−1.0	$-1.72^{+1.25}_{-0.91}$	$-1.43^{+1.13}_{-1.09}$	$-1.55^{+1.10}_{-1.00}$	$-1.49^{+1.00}_{-0.98}$
log τ	1.0	$0.18^{+1.31}_{-1.49}$	$0.50^{+1.09}_{-1.66}$	$0.61^{+0.98}_{-1.61}$	$0.79^{+0.87}_{-1.40}$
log f_c	−0.3	$-1.30^{+0.93}_{-1.13}$	$-1.31^{+0.94}_{-1.21}$	$-0.99^{+0.76}_{-1.27}$	$-0.76^{+0.54}_{-1.26}$

each of H_2O, O_3, and O_2. At this S/N, the oxygen mixing ratio is estimated to be above roughly 10^{-3}, indicating that we are unable to determine if O_2 is a major atmospheric constituent (i.e., present at the 1% level or more). Gravity (and, thus, planetary mass) remains undetected at all S/Ns, similar to the findings of Lupu et al. (2016). The surface albedo is unconstrained (or worse) at all S/Ns but shows a weak bias toward a higher value of $A_s \approx 0.3$ (log $A_s \approx -0.5$) at the highest S/Ns, which is likely due to the relatively large error bars at red wavelengths (driven primarily by low detector quantum efficiency) where we have the most sensitivity to the surface. We are able to get weak detections of τ and f_c, which are shown in Figure 10 to be correlated. Yet with these posteriors, we cannot rule out scenarios without cloud cover. We note that the drop-off in the posteriors of p_t and dp at higher pressure values likely result from the limiting conditions that the cloud base cannot extend below the surface pressure and the upper limit of the P_0 prior. The improved S/N leads to a posterior more concentrated around the true value for p_t, dp, and R_p. For improved constraints on cloud properties, it may be beneficial to observe time variability with photometry (e.g., Ford et al. 2001) or use polarimetry (e.g., Rossi & Stam 2017).

At a higher spectral resolution ($R = 140$), the improvement in detections and constraints begins at a lower S/N, as illustrated by Figure 12. Gravity remains undetected for all S/Ns. At an S/N equal to 5, P_0 and R_p have a weak detection and a detection, respectively. At S/N = 10, it is possible to detect H_2O, O_3, and O_2. As with the $R = 70$ case, the surface albedo is unconstrained (or worse) at all S/Ns, and, at the highest S/Ns, the model is biased toward $A_s \approx 0.3$ (as with $R = 70$). Moving to S/N = 15 adds a constraint to R_p, P_0, and O_3, as well as weak detections of cloud parameters. Increasing the S/N to 20 does not dramatically change the posterior distributions, although the posteriors for H_2O and O_2 become narrow enough to offer constraints. Here the constraint on O_2 suggests it is a major constituent in the atmosphere. In spite of the generous S/N, though, the 1σ uncertainties on the gas mixing ratios are not more precise than roughly an order of magnitude (see Table 5).

Considering both $R = 140$ and $R = 70$, we see that S/N = 5 data offer very little information about the planetary atmosphere. In the case of $R = 140$, S/N = 10 data offer detections but no constraints, and S/N = 20 data are required to constrain all included gas species. In other words, the conclusions we would draw about the planet (e.g., the amount

of gases, the bulk and cloud properties) improve significantly between S/N = 10 and 20. With $R = 70$, the boost from S/N = 10 to 15 provides weak detections of key atmospheric and surface parameters, and S/N = 20 data offer detections but few constraints (i.e., except on planetary radius).

For the *WFIRST* rendezvous data sets, we are able to infer very little information at an S/N of 5 or 10 except for weak detection of surface pressure and a detection of the planetary radius. All gases remain undetected at these S/Ns. The posterior distributions for most parameters do not vary much as S/N improves, although there are weak detections of cloud optical depth and fractional coverage at the highest S/Ns. Like all previous cases, we do not detect the surface gravity. At S/N = 15 and 20, the detection of f_c is unable to rule out scenarios with little cloud cover. To obtain weak detections of the atmospheric gases, we require an S/N of 20, but even here, the posteriors have tails that extend to near-zero mixing ratios.

To compare the performance of a *WFIRST* rendezvous scenario against *HabEx/LUVOIR* scenarios at $R = 70$ and 140, we plot the posterior distributions of the parameters for the S/N = 10 results from the *WFIRST* rendezvous and $R = 70$ and 140 in Figure 15. While this comparison sheds light on the corresponding trade-off in terms of parameter estimation for the same S/N, these cases do not represent equal integration times, which scale with resolution and S/N. If the dominant noise source does not depend on resolution (e.g., detector noise), the cases of $R = 140$ at S/N = 10, $R = 70$ at S/N = 20, and $R = 50$ at S/N = 28 would be roughly equal in integration time. However, if the dominant noise source does depend on resolution (e.g., exozodiacal dust), the cases of $R = 140$ at S/N = 10, $R = 70$ at S/N = 14, and $R = 50$ at S/N = 17 would have roughly equivalent integration times. Tables 7–9 allow approximate comparisons of these different scenarios, excluding a *WFIRST* rendezvous scenario at high S/N = 28 that we have not considered.

From Figure 15, we see that the performance of the *WFIRST* rendezvous retrieval is similar to that of $R = 70$ at S/N = 10. The noticeable difference is a weak detection of O_3 with $R = 70$. Because we adopt the photometric setup from *WFIRST* Design Cycle 7 through the shorter wavelengths, the data do not provide complete spectroscopic coverage across the significant O_3 feature from 0.5 to 0.7 μm, as in the case of *HabEx/LUVOIR* simulated data. Figure 5 shows the sampling of the forward-model spectrum for the three types of data sets we considered. We compare the spectral fits in Figure 14 and
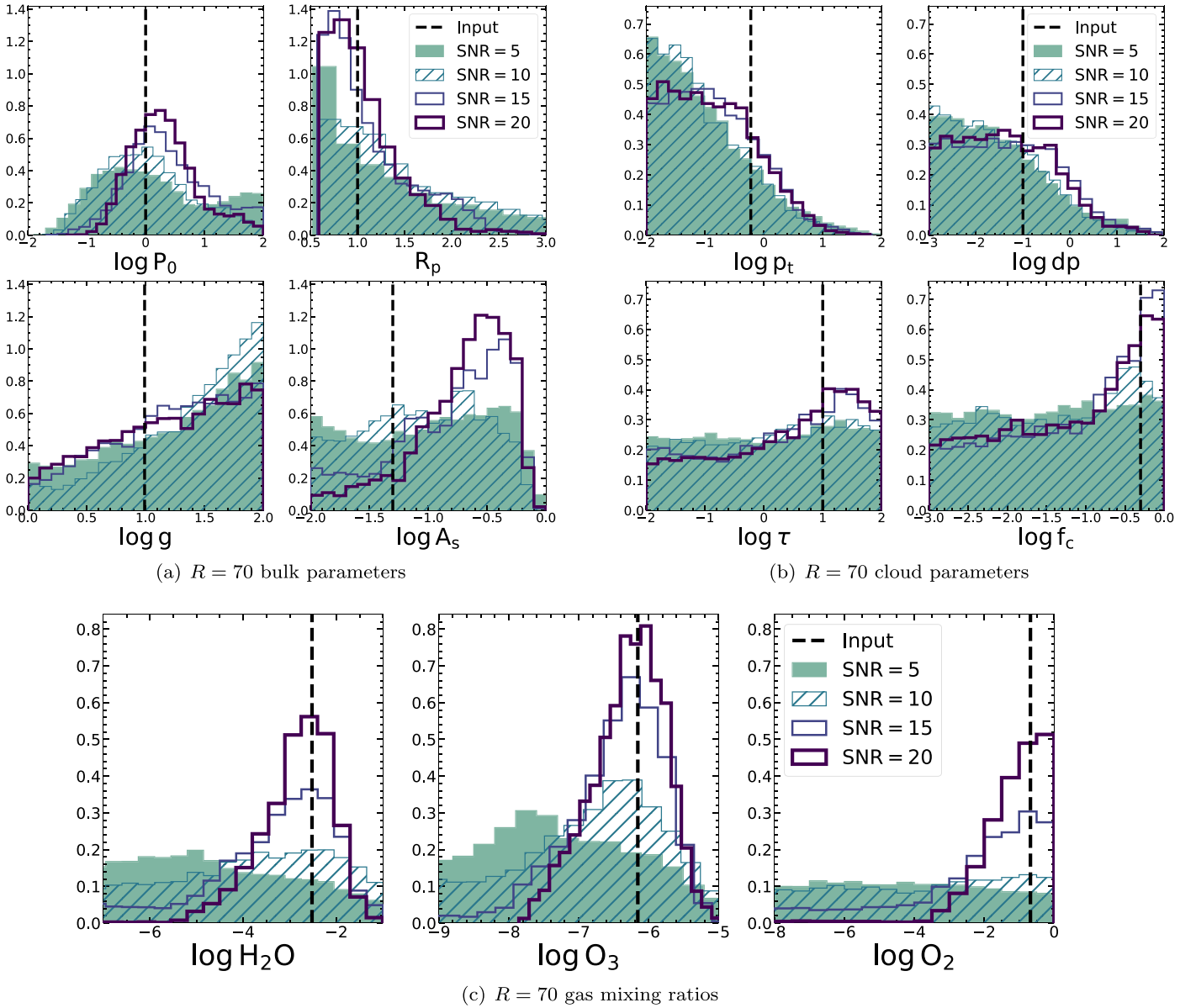
(a) $R = 70$ bulk parameters

(b) $R = 70$ cloud parameters

(c) $R = 70$ gas mixing ratios

Figure 11. Comparing 1D marginalized posterior distributions for all parameters for all S/N cases of $R = 70$. See Table 4 for the corresponding median retrieved value with uncertainties that indicates the 68% confidence interval. The overplotted dashed line represents the fiducial values from Table 1.

note the much wider spread in the possible fits for wavelengths shorter than 0.6 μm for the *WFIRST* rendezvous versus $R = 70$ or 140, which have continuous coverage in the full range. The $R = 140$, S/N = 10 data set was able to offer detections of all atmospheric gases, setting it apart from the other two. We stress, however, that constraints were only found at S/N = 20 and $R = 140$.

5.2. Considering Multiple Noise Instances

Our parameter estimations are likely to be optimistic as a consequence of our adoption of nonrandomized spectral data points in our faux observations. Thus, the requisite S/Ns for detection detailed above should be seen as lower limits. Ultimately, our decision to use nonrandomized data points stemmed from computational limitations (preventing us from running large numbers of randomized faux observations for each of our R–S/N pairs) and a desire to avoid the biases that can occur from attempting to make inferences from retrievals

performed on a single randomized faux observation (Lupu et al. 2016).

However, we deemed it necessary to investigate the consistency of our findings with respect to different noise instances. To work within our computational restrictions, we realized that cases such as $R = 70$ with S/N = 5 yielded little detection information for any parameter, even in the ideal scenario of nonrandomized data. We then decided to focus on two "threshold" cases based on the results from the nonrandomized data: $R = 140$ with S/N = 10 and $R = 70$ with S/N = 15. We ran 10 noise instances of these two cases where it is likely the optimistic nonrandomized data make the difference between detection and constraint for several parameters (see Tables 7 and 8).

Each noise instance is run for at least 10,000 steps in emcee. Figure 16 shows all the individual posteriors for the gas mixing ratios from each noise instance for $R = 70$, S/N = 15. We highlight the posteriors from one "outlier"
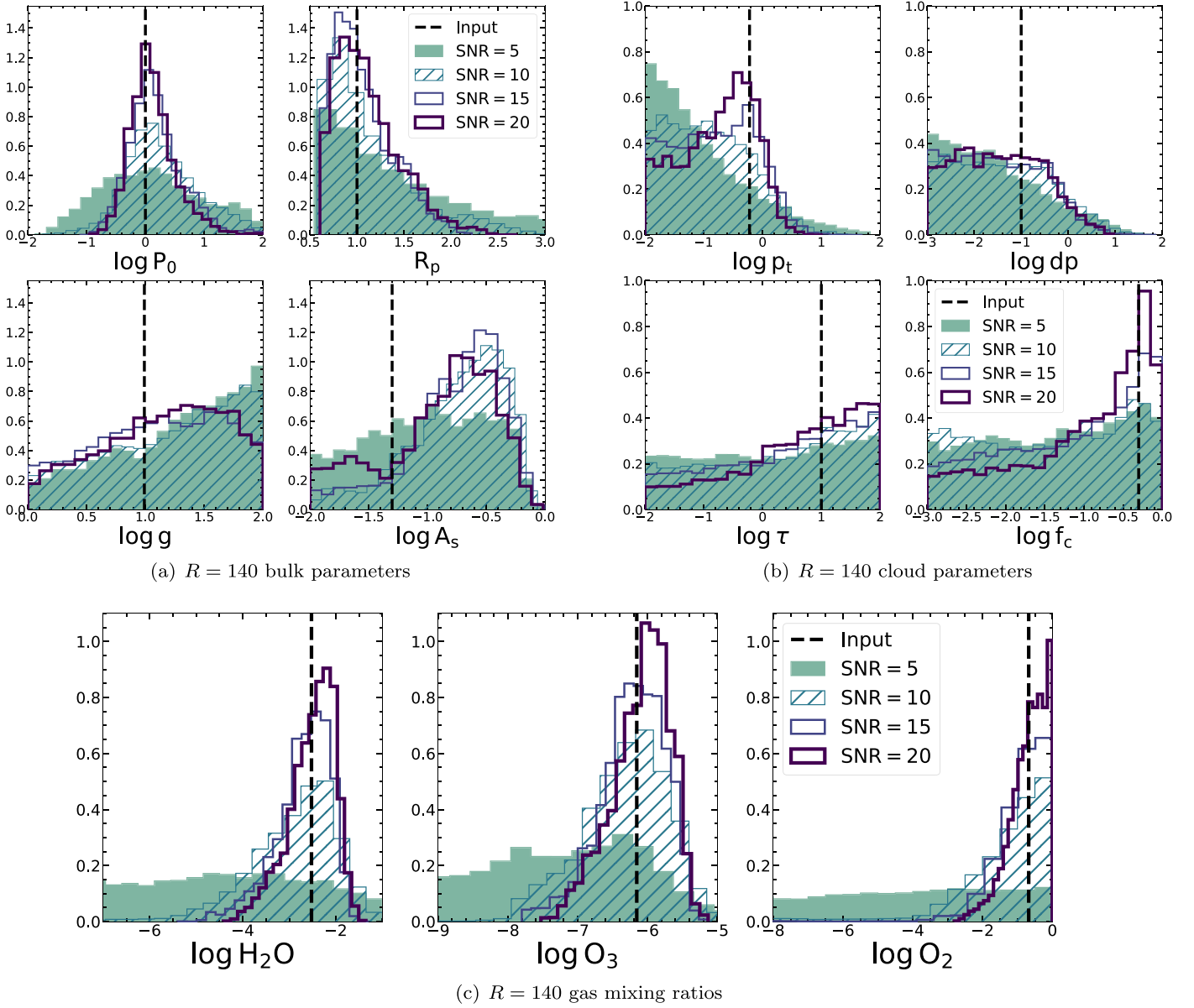
(a) $R = 140$ bulk parameters

(b) $R = 140$ cloud parameters

(c) $R = 140$ gas mixing ratios

Figure 12. Comparing 1D marginalized posterior distributions for all parameters for all S/N cases of $R = 140$. See Table 5 for the corresponding median retrieved value with uncertainties that indicates the 68% confidence interval. The overplotted dashed line represents the fiducial values from Table 1.

Table 6
WFIRST Rendezvous Retrieval Results, with Median Value and 1σ
Uncertainties of the Parameters

Parameter	Input	S/N = 5	S/N = 10	S/N = 15	S/N = 20
$\log H_2O$	-2.52	$-4.94^{+2.35}_{-2.05}$	$-4.89^{+2.48}_{-2.11}$	$-4.03^{+1.87}_{-2.52}$	$-3.11^{+1.17}_{-1.71}$
$\log P_0$	0.0	$-0.16^{+1.32}_{-0.80}$	$-0.19^{+1.03}_{-0.71}$	$0.03^{+1.16}_{-0.74}$	$0.45^{+1.01}_{-0.85}$
$\log O_3$	-6.15	$-7.66^{+1.65}_{-1.59}$	$-7.53^{+1.54}_{-1.66}$	$-7.16^{+1.19}_{-1.66}$	$-6.80^{+0.94}_{-1.30}$
$\log O_2$	-0.68	$-5.05^{+3.26}_{-3.39}$	$-4.89^{+3.43}_{-3.54}$	$-3.43^{+2.50}_{-4.41}$	$-2.26^{+1.71}_{-3.88}$
$\log P_0$	0.0	$-0.16^{+1.32}_{-0.80}$	$-0.19^{+1.03}_{-0.71}$	$0.03^{+1.16}_{-0.74}$	$0.45^{+1.01}_{-0.85}$
R_p	1.0	$1.13^{+1.60}_{-0.50}$	$1.13^{+1.27}_{-0.48}$	$1.02^{+1.10}_{-0.38}$	$0.80^{+0.81}_{-0.19}$
$\log g$	0.99	$1.42^{+0.42}_{-0.82}$	$1.45^{+0.41}_{-0.75}$	$1.41^{+0.43}_{-0.83}$	$1.26^{+0.52}_{-0.83}$
$\log A_s$	-1.3	$-0.89^{+0.63}_{-0.74}$	$-0.84^{+0.56}_{-0.70}$	$-0.95^{+0.64}_{-0.68}$	$-0.76^{+0.53}_{-0.79}$
$\log p_t$	-0.22	$-1.26^{+0.98}_{-0.52}$	$-1.24^{+0.83}_{-0.55}$	$-1.23^{+0.84}_{-0.57}$	$-0.84^{+0.89}_{-0.73}$
$\log dp$	-1.0	$-1.75^{+1.16}_{-0.87}$	$-1.70^{+0.94}_{-0.87}$	$-1.46^{+1.14}_{-1.06}$	$-1.49^{+1.49}_{-1.03}$
$\log \tau$	1.0	$0.03^{+1.33}_{-1.39}$	$0.05^{+1.42}_{-1.40}$	$0.61^{+0.94}_{-1.55}$	$0.99^{+0.73}_{-1.44}$
$\log f_c$	-0.3	$-1.41^{+0.96}_{-1.07}$	$-1.42^{+1.00}_{-1.07}$	$-0.82^{+0.60}_{-1.28}$	$-0.58^{+0.41}_{-0.97}$

case where there is no oxygen detection. The corresponding set of data points is shown as well. This highlights the fact that single noise instances can mislead our interpretation and the benefit of having many noise instances run to obtain a more comprehensive understanding of the state of an atmosphere.

To summarize the noise instance results, we concatenate samples from the last 1000 steps in each noise instance and construct an averaged set of posteriors. We are able to do this because the noise instances are equally likely, having been drawn in the same manner from a Gaussian with set parameters (i.e., the same S/N as the standard deviation). In Figure 17, we plot the combined posteriors of the 10 noise instances of $R = 70$, S/N = 15 and compare them to the posterior from the last 5000 steps of the nonrandomized data case. We illustrate the same comparison for $R = 140$, S/N = 10 in Figure 18. We overplot the truth values, as well as the 68% confidence interval and median value, for each parameter from the combined noise-instances posterior and the nonrandomized data posterior.
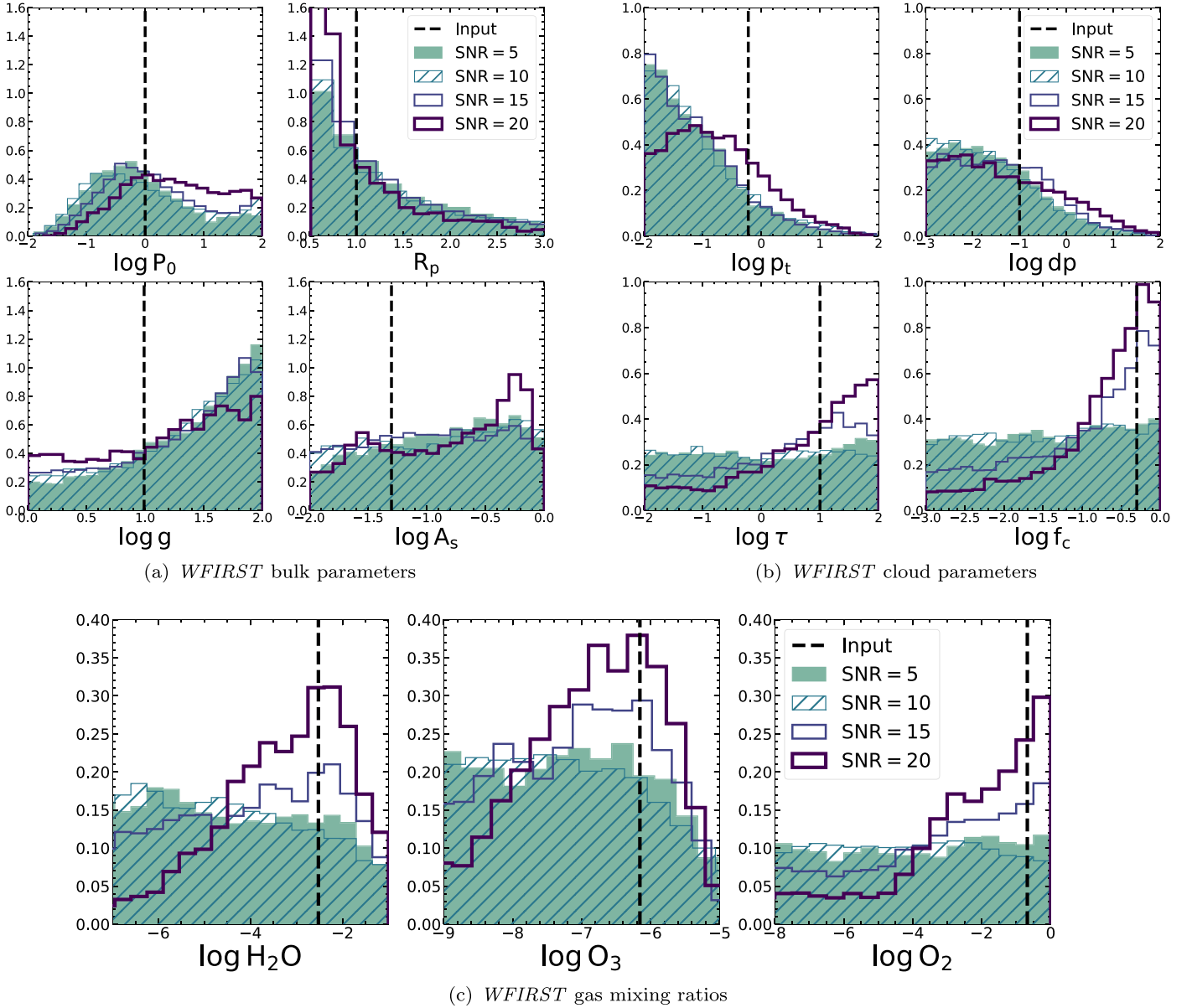
Figure 13. Comparing 1D marginalized posterior distributions for all parameters for all S/N cases of a *WFIRST* rendezvous scenario. See Table 6 for the corresponding median retrieved value with uncertainties that indicates the 68% confidence interval. The overplotted dashed line represents the fiducial values from Table 1.

For all parameters in both the $R = 70$ and $R = 140$ cases, we find that the average posterior from the 10 noise instances agrees with the posterior from the nonrandomized data set qualitatively. Their medians and 68% confidence interval ranges are also similar with significant overlap. The overall conclusions we can draw from the average posteriors do not appear to differ much from those using the nonrandomized data set posteriors.

5.3. Implications for Future Direct-imaging Missions

Future space-based direct-imaging missions will have a diversity of goals for exoplanet studies and will likely emphasize the detection and characterization of Earth-like exoplanets. For the detection of oxygen and ozone—which are key biosignature gases—in the atmospheres of Earth twins, our results indicate that spectra at a minimum characteristic S/N of

10 will suffice if at $R = 140$, while data at an S/N of at least 15–20 would be needed at $R = 70$. For a *WFIRST* rendezvous-like observing setup, these gases would only be weakly detected, even at an S/N of 20. Methane, which is another important biosignature gas, has no strong signatures in the visible wavelength range for the modern Earth, so we did not consider detection of this gas. Thus, we could not use our simulated data and retrievals to argue for detections of atmospheric chemical disequilibrium (Sagan et al. 1993; Krissansen-Totton et al. 2016).

Key habitability indicators include atmospheric water vapor and surface pressure. Detecting the former requires an S/N of 15–20 at $R = 70$ but only an S/N of 10 at $R = 140$. Surface pressure can be constrained to within an order of magnitude for S/N $\gtrsim 15$ at $R = 140$, although the overall lack of temperature information in these reflected-light spectra would make it impossible to use pressure/temperature data to argue for
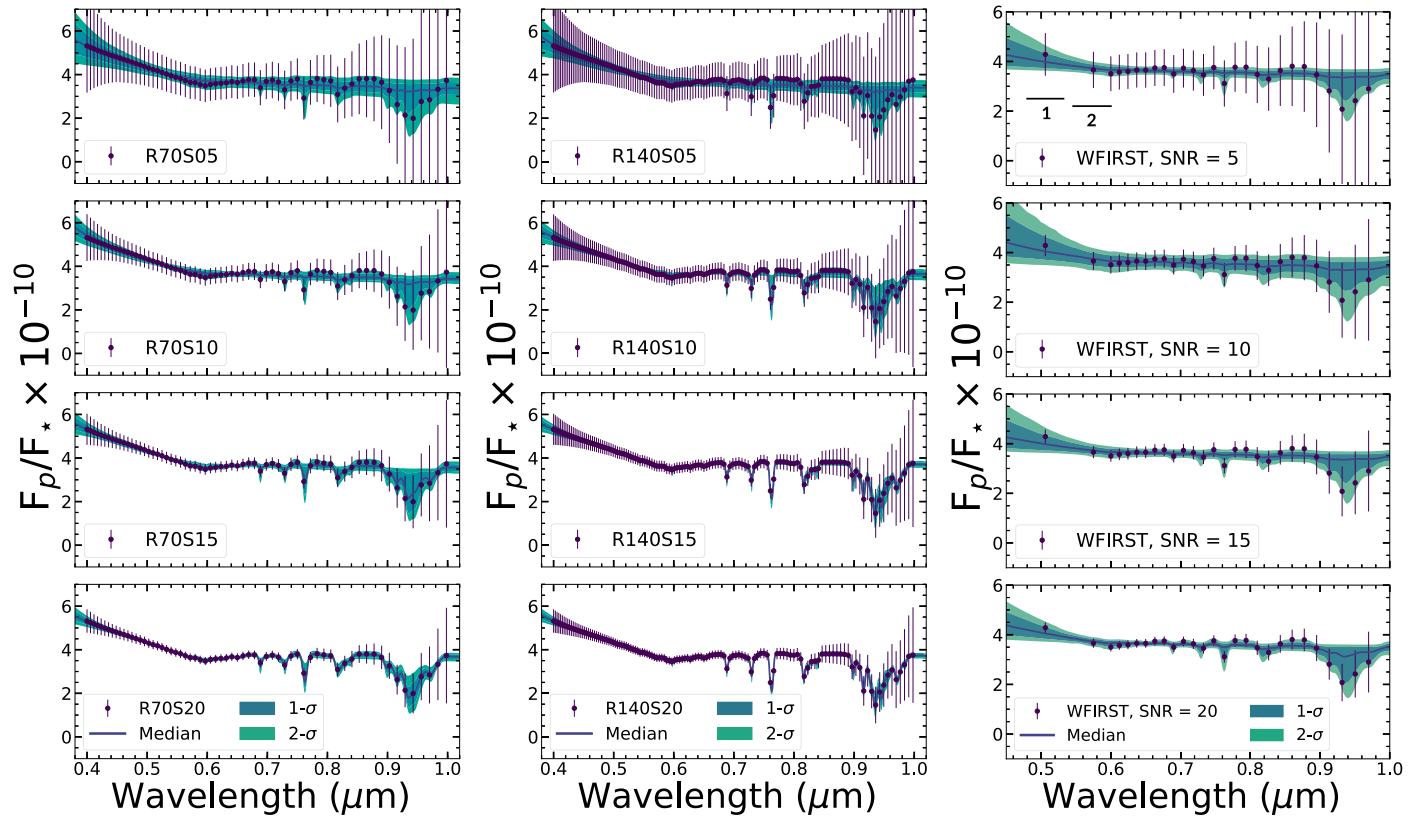
Figure 14. Spectra generated with 1000 randomly drawn sets of parameters sampled with the retrievals plotted with (left) $R = 70$ data for S/N = 5, 10, 15, and 20; (middle) $R = 140$ data for S/N = 5, 10, 15, and 20; and (right) *WFIRST* rendezvous data at S/N = 5, 10, 15, and 20. Here "1" and "2" mark the span of the *WFIRST* Design Cycle 7 filters (see Table 2). Lighter contours (green) represent 2σ fits, while darker contours (blue) represent 1σ fits. The solid line represents the median fit.

Table 7
$R = 70$: Strength of Detection for a Set of Key Parameters as a Function of S/N

Parameter	S/N = 5	S/N = 10	S/N = 15	S/N = 20
H_2O	–	–	W	D
O_3	–	W	W	D
O_2	–	–	W	D
P_0	W	W	W	D
R_p	D	D	D	C

Note. Weak detection ("W") corresponds to a posterior distribution with a marked peak but also a substantial tail toward extreme values. Detection ("D") refers to a peaked posterior without tails toward extreme values but with a 1σ width larger than an order of magnitude. Constraint ("C") is defined as a peaked posterior distribution with a 1σ width less than an order of magnitude. Nondetections, or flat posteriors across the entire (or nearly entire) prior range, are marked with "–."

habitability (Robinson 2017). Surface temperature information may then need to come from climate-modeling investigations that are constrained by retrieved gas mixing ratios.

For all of our observing setups, the data yield detections of, and in some cases constraints on, the planetary radius. Except at an S/N of 20 for $R = 70$ or S/N > 15 for $R = 140$, the posterior distributions are not well enough constrained to distinguish an Earth/super-Earth ($R_p < 1.5\,R_\oplus$) from a mini-Neptune based on size alone, although the data do rule out planetary sizes larger than Neptune. Additional atmospheric information (e.g., composition) could potentially be used to help distinguish between terrestrial planets and mini-Neptunes.

These findings are consistent with the gas giant–focused work of Nayak et al. (2017), who noted that observations at multiple phase angles can also help to better constrain planetary size. Our overall lack of surface gravity constraints, paired with the weak constraints on planet size, implies that we do not have a constraint on the planetary mass. Follow-up (or precursor) radial velocity observations (or, potentially, astrometric observations) could offer additional constraints on planet mass.

We can make rough comparisons of our R–S/N results to those of Brandt & Spiegel (2014), who used minimally parametric models to investigate detections of O_2 and H_2O for Earth twins. These comparisons are not direct, however, as Brandt & Spiegel (2014) were fitting for fewer parameters (8 versus our 11) and also only assumed that the S/N was proportional to planetary reflectance (versus our more complicated scaling, as shown in Figure 6). For O_2, Brandt & Spiegel (2014) found $R = 150$ and S/N = 6 for a 90% detection probability, which is consistent with our $R = 140$ posteriors moving from a nondetection at S/N = 5 to a detection at S/N = 10. When investigating H_2O, Brandt & Spiegel (2014) found $R = 40$ and S/N = 7.5 or $R = 150$ and S/N = 3.3 for a 90% detection probability. Using Figure 6 to scale our S/Ns to 890 nm (i.e., to the continuum just shortward of the 950 nm water vapor band), at $R = 50$, we only find a weak detection of H_2O for S/N$_{890nm}$ = 10, and at $R = 140$, we transition from a water vapor nondetection to a detection between an S/N$_{890nm}$ of 2.5–5. Taken all together, these comparisons indicate that we agree with Brandt & Spiegel (2014) at higher spectral resolution ($R = 140$–150) but that detection of H_2O at lower
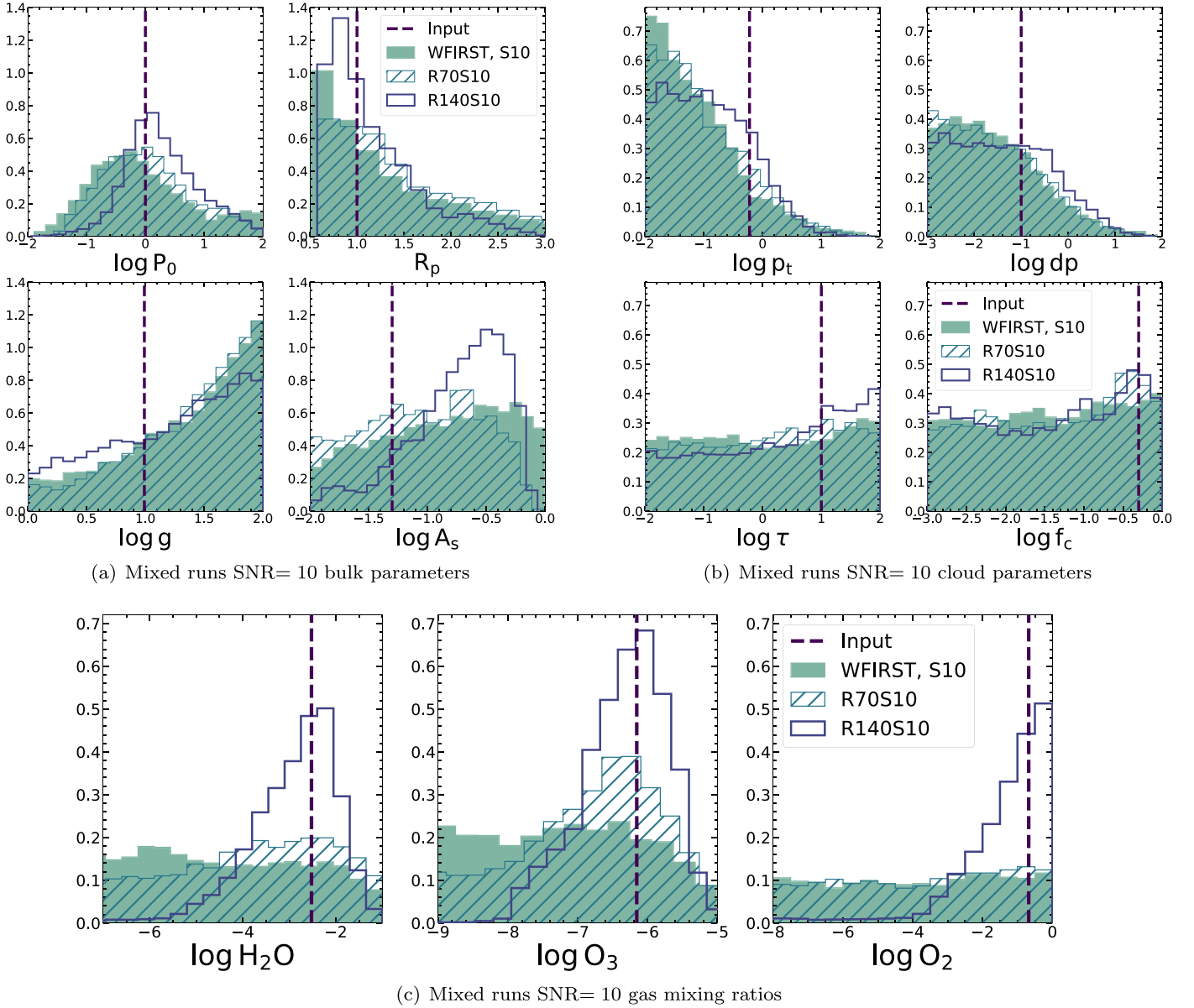
(a) Mixed runs SNR= 10 bulk parameters

(b) Mixed runs SNR= 10 cloud parameters

(c) Mixed runs SNR= 10 gas mixing ratios

Figure 15. Comparing the posteriors for all parameters for S/N = 10 cases of the *WFIRST* rendezvous and $R = 70$ and 140. The overplotted dashed line represents the fiducial values from Table 1.

Table 8
$R = 140$: Strength of Detection for a Set of Key Parameters as a Function of S/N

Parameter	S/N = 5	S/N = 10	S/N = 15	S/N = 20
H_2O	–	D	D	C
O_3	–	D	C	C
O_2	–	D	D	C
P_0	W	D	C	C
R_p	D	D	C	C

Note. Weak detection ("W") corresponds to a posterior distribution with a marked peak but also a substantial tail toward extreme values. Detection ("D") refers to a peaked posterior without tails toward extreme values but with a 1σ width larger than an order of magnitude. Constraint ("C") is defined as a peaked posterior distribution with a 1σ width less than an order of magnitude. Nondetections, or flat posteriors across the entire (or nearly entire) prior range, are marked with "–."

spectral resolution ($R = 50$) will likely require higher S/Ns than originally indicated.

The discussion above emphasizes mere detections, not constraints (which, again, we define as having peaked posterior distributions with 1σ widths less than an order of magnitude). While uncertain, we anticipate that characterization of climate, habitability, and life likely require constraints, not simple detections. Here, as is shown in Table 8, only $R = 140$ and S/N = 20 observations offer the appropriate constraints. Thus, future space-based high-contrast imaging missions with goals of characterizing Earth-like planetary environments are likely to need to achieve $R = 140$ and S/N = 20 observations (or better). Of course, combining near-infrared capabilities, which would provide access to additional gas absorption bands, may help loosen these requirements.
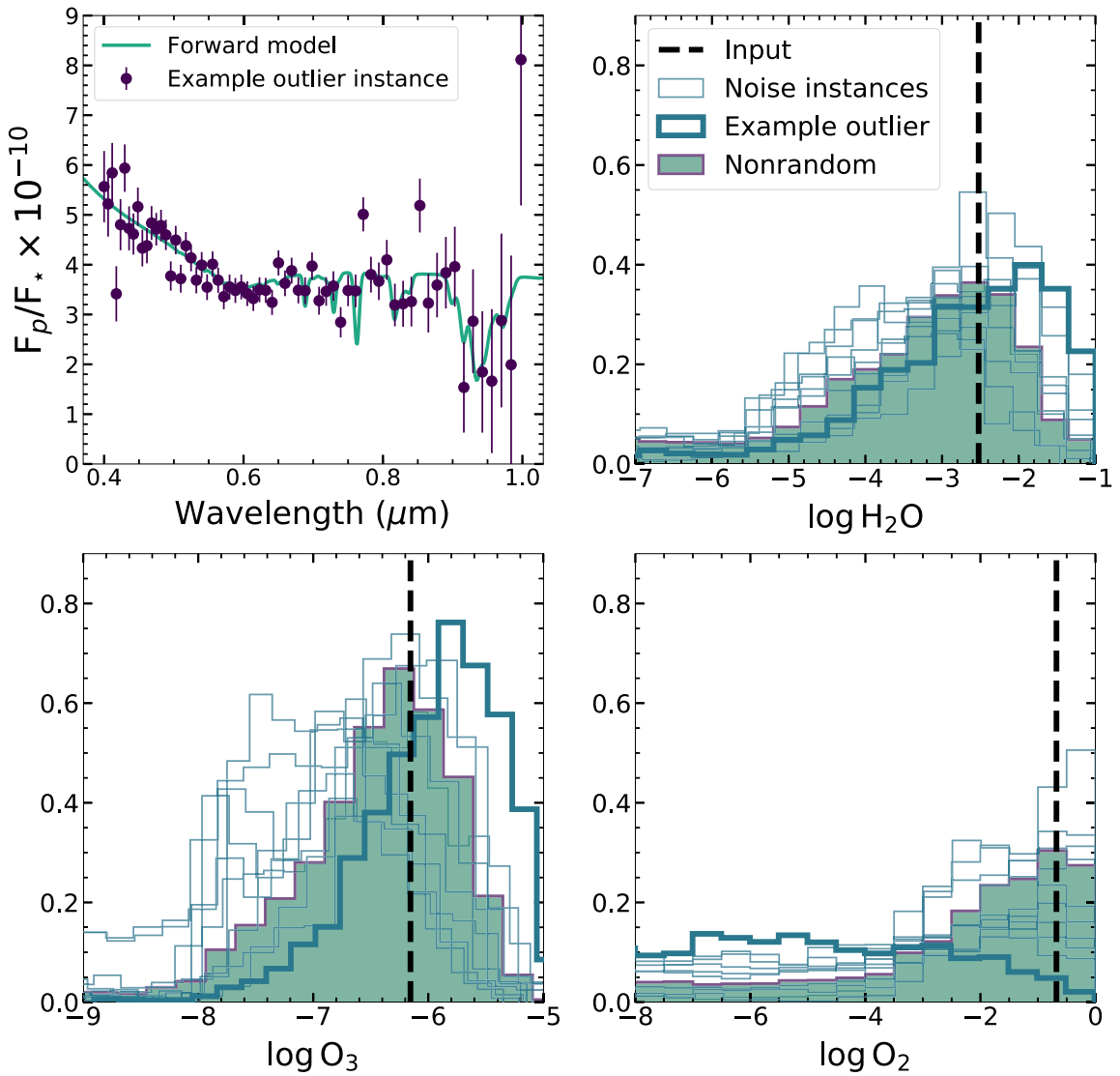
Figure 16. The top left panel shows one of the 10 noise instances we retrieved on for $R = 70$, S/N = 15 data, plotted along with the forward-model spectrum at $R \sim 70$. The remaining three panels show the gas mixing ratio posteriors (H_2O, O_3, O_2) of all of the 10 noise instances of $R = 70$, S/N = 15. In addition, we show the corresponding posterior distributions from the nonrandomized data set (seen originally in Figure 11) for comparison. The set of posteriors that corresponds to the noise instance in the top left panel is the set of bolded distributions. The vertical dashed lines represent the input values of the parameters.

5.4. Impacts of Model Assumptions

Several key assumptions adopted in this study warrant further comment. First, as noted earlier, we do not retrieve on planetary phase angle and planet–star distance, both of which influence the planet-to-star flux ratio. Thus, in effect, we are assuming that the planetary system has been revisited multiple times for photometric and astrometric measurements, such that the planetary orbit is reasonably well-constrained (i.e., that the orbital distance and phase angle are not the dominant sources of uncertainty when interpreting the observed planet-to-star flux ratio spectrum). If the orbit is not well-constrained, Nayak et al. (2017) showed that strong correlations can exist between the retrieved phase angle and the planet radius.

Second, we have assumed detector-dominated noise and a quantum efficiency appropriate for the *WFIRST*/CGI for all of our observational setups. While this is likely a fair assumption for our *WFIRST* rendezvous studies, it is likely that detector development will lead to major improvements in instrumentation for a *HabEx*/*LUVOIR*-like mission. Here the rapid

decrease into the red due to detector quantum efficiency may not be as dramatic, implying that spectra would have relatively more information content at red wavelengths as compared to the present study. Furthermore, a *HabEx*/*LUVOIR*-like mission may no longer be in the detector-dominated noise regime. In the limit of noise dominated by astrophysical sources (e.g., exozodiacal light or stellar leakage), the S/N only varies as $\sqrt{qTB_\lambda}$.

Finally, we adopt a relatively simple parameterization of cloud 3D structure. Specifically, we allow for only a single cloud deck in the atmosphere, and we then permit these clouds to have some fractional coverage over the entire planet. This parameterization of fractional cloudiness implies uniform latitudinal and longitudinal distribution of patchy clouds. In reality, clouds on Earth have a complex distribution in altitude, latitude, and longitude (Stubenrauch et al. 2013), and variations in time also have an observational impact (Cowan et al. 2009; Cowan & Fujii 2017). However, given the overall inability of our retrievals to constrain cloud parameters (at least at the S/Ns
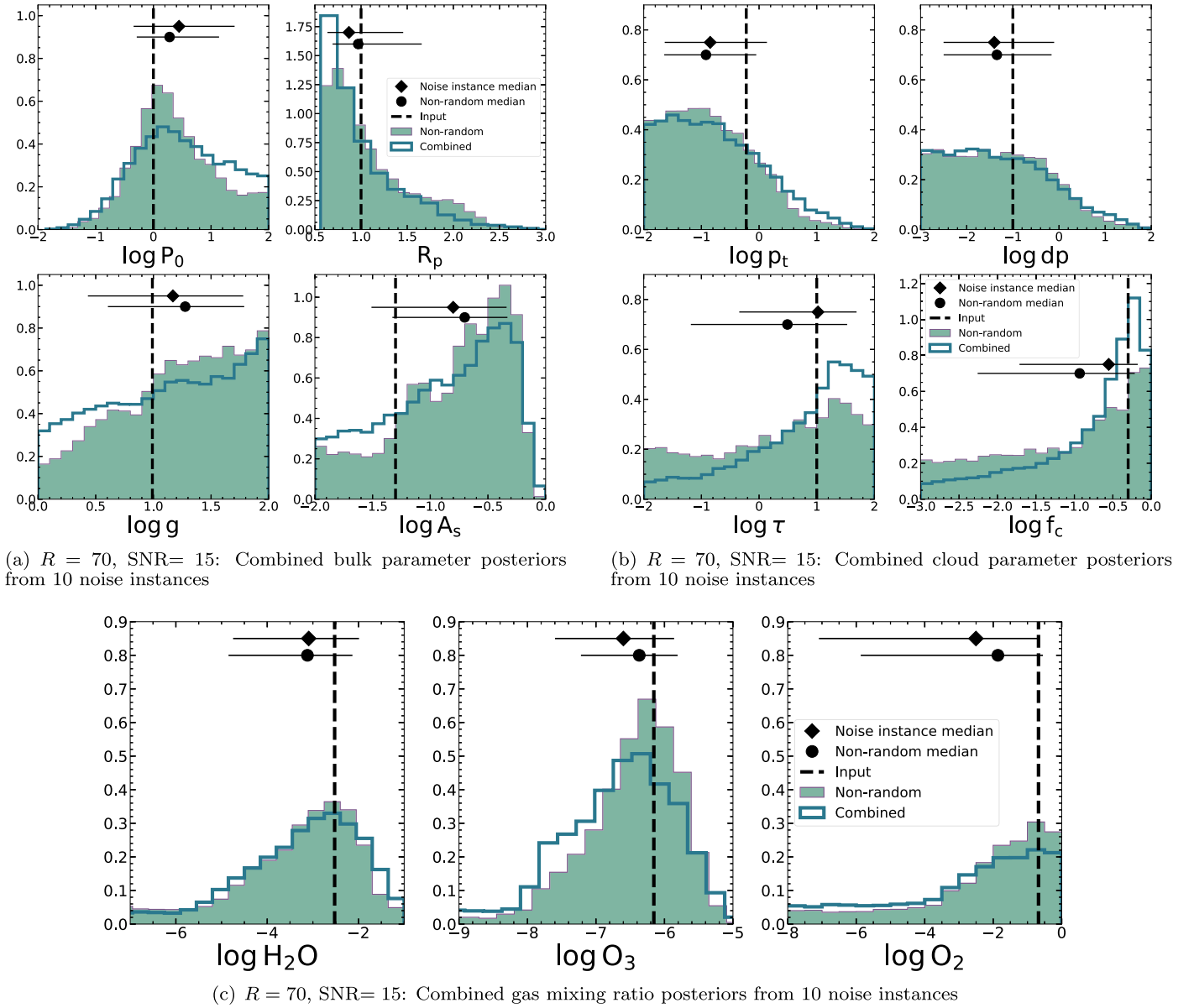
(a) $R = 70$, SNR= 15: Combined bulk parameter posteriors from 10 noise instances

(b) $R = 70$, SNR= 15: Combined cloud parameter posteriors from 10 noise instances

(c) $R = 70$, SNR= 15: Combined gas mixing ratio posteriors from 10 noise instances

Figure 17. Combined posterior distributions from 10 noise instances of $R = 70$, S/N = 15 compared to the posteriors from the nonrandomized data set (see also Figure 11). The diamond represents the median value of each combined posterior, while the circle is the median of the nonrandomized data set posterior. Each median is plotted along with the 68% confidence interval from the same distribution. The vertical dashed lines represent the input values of the parameters.

investigated here; see also Lupu et al. 2016; Nayak et al. 2017), it seems challenging for future space-based exoplanet characterization missions to detect (or constrain) more complex cloud distributions with the types of observations studied here and data of similar quality.

5.5. Future Work

Our current forward model is able to include both CO_2 and CH_4, although we did not retrieve on these gases in the current study due to their overall lack of strong features in the visible wavelength range for modern Earth. However, these species do have stronger features in the near-infrared wavelength range. As both of the *HabEx* and *LUVOIR* concepts are considering near-infrared capabilities, it will be essential to extend our current studies to longer wavelengths and investigate whether or not constraints on additional gases (i.e., beyond water, oxygen, and ozone) can be achieved at these wavelengths.

Additionally, given the likely huge diversity of exoplanets that will be discovered by future missions (and that have already been identified and studied by *Kepler*, *Hubble*, and *Spitzer*), it will be necessary to extend our parameter estimation studies to include a wider range of worlds. Both super-Earths and mini-Neptunes are more favorable targets for a *WFIRST* rendezvous mission and may also be easier targets for *HabEx*/ *LUVOIR*-like missions. Our forward model is already capable of simulating these types of worlds, and we anticipate emphasizing a variety of exoplanet types in future studies. Such future studies may also include retrievals on planetary phase angle, which would be relevant to observing scenarios where the planetary orbit is poorly constrained.

6. Summary

We have developed a retrieval framework for constraining atmospheric properties of an Earth-like exoplanet observed
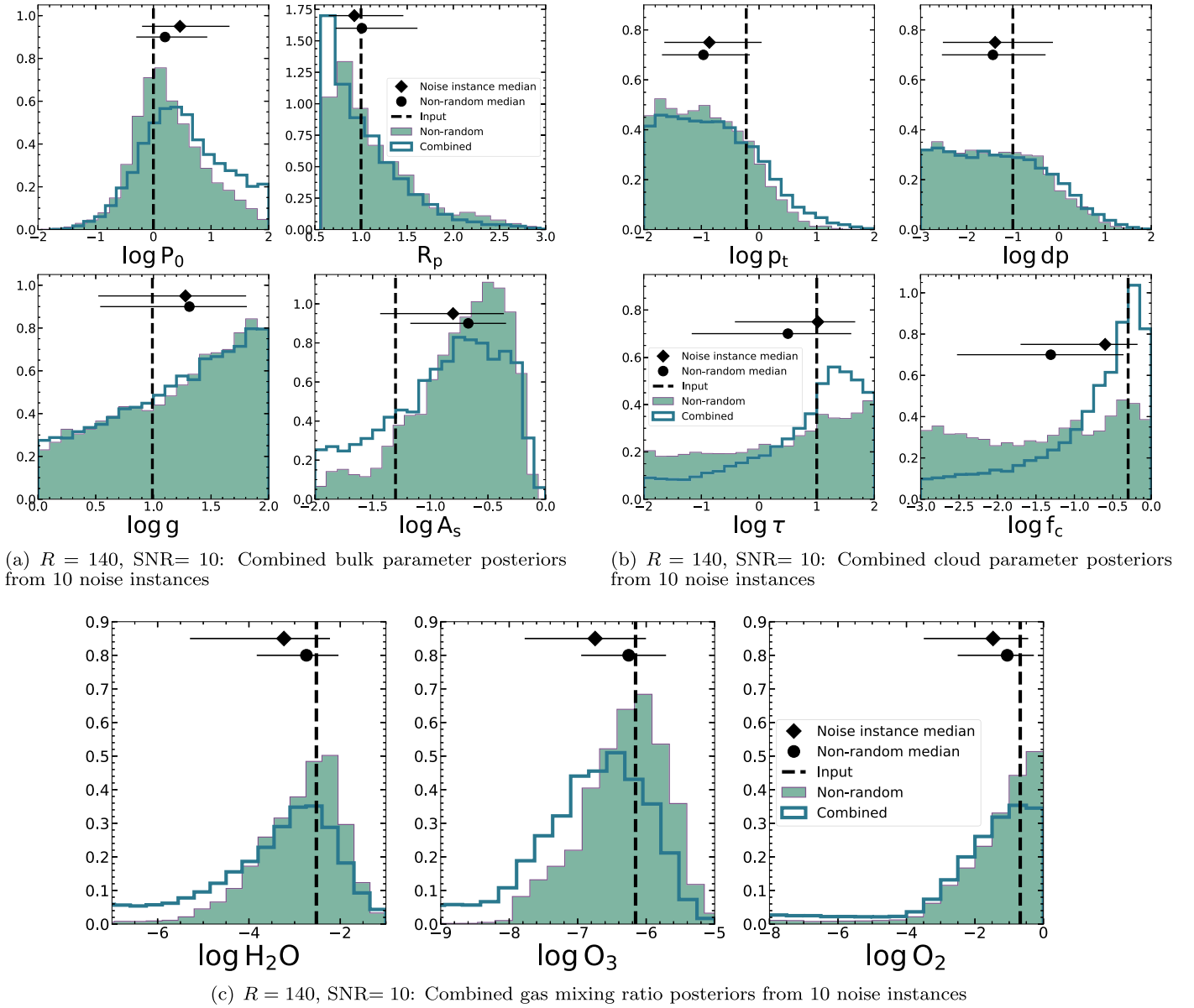
(a) $R = 140$, SNR$= 10$: Combined bulk parameter posteriors from 10 noise instances

(b) $R = 140$, SNR$= 10$: Combined cloud parameter posteriors from 10 noise instances

(c) $R = 140$, SNR$= 10$: Combined gas mixing ratio posteriors from 10 noise instances

Figure 18. Combined posterior distributions from 10 noise instances of $R = 140$, S/N $= 10$ compared to the posteriors from the nonrandomized data set (see also Figure 12). The diamond represents the median value of each combined posterior, while the circle is the median of the nonrandomized data set posterior. Each median is plotted along with the 68% confidence interval from the same distribution. The vertical dashed lines represent the input values of the parameters.

with reflected-light spectroscopy spanning the visible range (0.4–1.0 μm). We have upgraded an existing, well-tested albedo model to generate high-resolution geometric albedo spectra used to simulate data at resolutions and quality relevant to future telescopes, such as the *HabEx* and *LUVOIR* mission concepts. We combined our albedo model with Bayesian inference techniques and applied MCMC sampling to perform parameter estimation. The data we considered were for *WFIRST* paired with a starshade (i.e., the rendezvous scenario), $R = 70$, and $R = 140$ at S/N $= 5$, 10, 15, and 20. We validated our forward model, and we demonstrated the successful application of our retrieval approach by gradually adding complexity to our inverse analyses.

Following work by Lupu et al. (2016) and Nayak et al. (2017), who constructed a retrieval framework for gas giants in reflected light, we made several modifications to the albedo model featured in these previous studies. Our model has a

reflective surface; absorption due to water vapor, oxygen, and ozone; Rayleigh scattering from nitrogen and other key gases; pressure-dependent opacities; an adaptive pressure grid; and a single-layer water vapor cloud layer with fractional cloudiness. We performed our retrievals with the goal of estimating our ability to detect and constrain the atmosphere of an Earth twin. We found that $R = 70$, S/N $= 15$ data allowed us to weakly detect surface pressure, as well as water vapor, ozone, and oxygen. At $R = 140$, we found that S/N $= 10$ was needed to more firmly detect these parameters. At $R = 140$, an S/N of 20 was needed to constrain key planetary parameters, and $R = 70$ data at this S/N offered extremely few constraints. A *WFIRST* rendezvous scenario, with its photometric points and lower-resolution spectrum ($R = 50$), is only able to offer limited diagnostic information. For example, at S/N $= 10$, we only weakly detect and detect surface pressure and planetary radius, respectively. To weakly detect the gases, *WFIRST* rendezvous

Table 9
WFIRST: Strength of Detection for a Set of Key Parameters as a Function of S/N

Parameter	S/N = 5	S/N = 10	S/N = 15	S/N = 20
H_2O	–	–	–	W
O_3	–	–	–	W
O_2	–	–	W	W
P_0	W	W	W	W
R_p	D	D	D	D

Note. Weak detection ("W") corresponds to a posterior distribution with a marked peak but also a substantial tail toward extreme values. Detection ("D") refers to a peaked posterior without tails toward extreme values but with a 1σ width larger than an order of magnitude. Constraint ("C") is defined as a peaked posterior distribution with a 1σ width less than an order of magnitude. Nondetections, or flat posteriors across the entire (or nearly entire) prior range, are marked with "–."

data needed to be at least S/N = 20. Throughout our runs, we find that we are unable to accurately constrain surface albedo or place estimates on the surface gravity, although we can straightforwardly rule out planetary sizes above roughly the radius of Uranus or Neptune.

Our findings demonstrate that direct imaging of Earth-like exoplanets in reflected light offers a promising path forward for detecting and constraining atmospheric biosignature gases. Instrument spectral resolution for future missions strongly impacts the requisite S/Ns for detection and characterization, and this must be taken into consideration during mission design. Thus, the scientific yield of future space-based exoplanet direct-imaging missions can only be maximized by simultaneously considering mission characterization goals, integration time constraints, and instrument spectral performance.

ORCID iDs

Y. Katherina Feng ⓘ https://orcid.org/0000-0002-5032-5060
Tyler D. Robinson ⓘ https://orcid.org/0000-0002-3196-414X
Jonathan J. Fortney ⓘ https://orcid.org/0000-0002-9843-4354
Roxana E. Lupu ⓘ https://orcid.org/0000-0003-3444-5908
Mark S. Marley ⓘ https://orcid.org/0000-0002-5251-2943
Nikole K. Lewis ⓘ https://orcid.org/0000-0002-8507-1304
Bruce Macintosh ⓘ https://orcid.org/0000-0003-1212-7538
Michael R. Line ⓘ https://orcid.org/0000-0002-2338-476X

References

Allen, C., & Cox, A. 2000, Allen's Astrophysical Quantities, Allenas Astrophysical Quantities (Berlin: Springer)
Balasubramanian, K., White, V., Yee, K., et al. 2016, JATIS, 2, 011005
Barman, T. S., Konopacky, Q. M., Macintosh, B., & Marois, C. 2015, ApJ, 804, 61
Barman, T. S., Macintosh, B., Konopacky, Q. M., & Marois, C. 2011, ApJ, 733, 65
Benneke, B., & Seager, S. 2012, ApJ, 753, 100
Borucki, W. J., Koch, D. G., Basri, G., et al. 2011, ApJ, 728, 117
Borucki, W. J., Koch, D. G., Lissauer, J. J., et al. 2003, Proc. SPIE, 4854, 129
Brandt, T. D., & Spiegel, D. S. 2014, PNAS, 111, 13278
Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8
Burrows, A. 2014, arXiv:1412.6097
Cady, E., Prada, C. M., An, X., et al. 2016, JATIS, 2, 011004
Cahoy, K. L., Marley, M. S., & Fortney, J. J. 2010, ApJ, 724, 189
Cash, W. 2006, Natur, 442, 51
Charbonneau, D., Brown, T. M., Noyes, R. W., & Gilliland, R. L. 2002, ApJ, 568, 377
Chen, J., & Kipping, D. 2017, ApJ, 834, 17
Cowan, N. B., Agol, E., Meadows, V. S., et al. 2009, ApJ, 700, 915
Cowan, N. B., & Fujii, Y. 2017, arXiv:1704.07832
Crill, B., & Siegler, N. 2017, arXiv:1709.06660
de Wit, J., Wakeford, H. R., Gillon, M., et al. 2016, Natur, 537, 69
Ford, E. B., Seager, S., & Turner, E. L. 2001, Natur, 412, 885
Foreman-Mackey, D. 2016, JOSS, 1, 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Goodman, J., & Weare, J. 2010, Commun. Appl. Math. Comput. Sci., 5, 65
Greco, J. P., & Burrows, A. 2015, ApJ, 808, 172
Groff, T., Gong, Q., Mandell, A. M., et al. 2018, in AAS Meeting 231 Abstracts, 355.47
Hansen, J. E., & Travis, L. D. 1974, SSRv, 16, 527
Horak, H. G. 1950, ApJ, 112, 445
Horak, H. G., & Little, S. J. 1965, ApJS, 11, 373
Hunter, J. D. 2007, CSE, 9, 90
Irwin, P. G. J., Teanby, N. A., de Kok, R., et al. 2008, JQSRT, 109, 1136
Kasdin, N. J., Lisman, D., Shaklan, S., et al. 2012, Proc. SPIE, 8442, 84420A
Kasting, J. F., Whitmire, D. P., & Reynolds, R. T. 1993, Icar, 101, 108
Kattawar, G. W. 1975, JQSRT, 15, 839
Konopacky, Q. M., Barman, T. S., Macintosh, B. A., & Marois, C. 2013, Sci, 339, 1398
Krissansen-Totton, J., Bergsman, D. S., & Catling, D. C. 2016, AsBio, 16, 39
Krist, J., Nemati, B., & Mennesson, B. 2016, JATIS, 2, 011003
Lee, J.-M., Fletcher, L. N., & Irwin, P. G. J. 2012, MNRAS, 420, 170
Line, M. R., Wolf, A. S., Zhang, X., et al. 2013, ApJ, 775, 137
Lupu, R. E., Marley, M. S., Lewis, N., et al. 2016, AJ, 152, 217
Macintosh, B., Graham, J. R., Barman, T., et al. 2015, Sci, 350, 64
Madhusudhan, N., & Seager, S. 2009, ApJ, 707, 24
Marley, M., Lupu, R., Lewis, N., et al. 2014, arXiv:1412.8440
Marley, M. S., Gelino, C., Stephens, D., Lunine, J. I., & Freedman, R. 1999, ApJ, 513, 879
Marley, M. S., & Robinson, T. D. 2015, ARA&A, 53, 279
Mawet, D., Ruane, G., Xuan, W., et al. 2017, ApJ, 838, 92
Mayor, M., & Queloz, D. 1995, Natur, 378, 355
McClatchey, R. A., Fenn, R. W., Selby, J. E. A., Volz, F. E., & Garing, J. S. 1972, Optical Properties of the Atmosphere (3rd ed.; Wright-Patterson Air Force Base, OH: Air Force Cambridge Research Laboratories)
McKay, C. P., Pollack, J. B., & Courtin, R. 1989, Icar, 80, 23
Meador, W. E., & Weaver, W. R. 1980, JAtS, 37, 630
Meadows, V. S., & Crisp, D. 1996, JGR, 101, 4595

Mennesson, B., Gaudi, S., Seager, S., et al. 2016, Proc. SPIE, 9904, 99040L

Nayak, M., Lupu, R., Marley, M. S., et al. 2017, PASP, 129, 034401

Noecker, M. C., Zhao, F., Demers, R., et al. 2016, JATIS, 2, 011001

Peterson, B. M., Fischer, D. & LUVOIR Science and Technology Definition Team 2017, AAS Meeting 229 Abstracts, 405.04

Peterson, P. 2009, JCSE, 4, 296

Robinson, T. D. 2017, arXiv:1701.05205

Robinson, T. D., Ennico, K., Meadows, V. S., et al. 2014, ApJ, 787, 171

Robinson, T. D., Meadows, V. S., Crisp, D., et al. 2011, AsBio, 11, 393

Robinson, T. D., Stapelfeldt, K. R., & Marley, M. S. 2016, PASP, 128, 025003

Rodgers, C. D. 1976, RvGSP, 14, 609

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

Rossi, L., & Stam, D. M. 2017, A&A, 607, A57

Rothman, L. S., Gordon, I. E., Babikov, Y., et al. 2013, JQSRT, 130, 4

Sagan, C., Thompson, W. R., Carlson, R., Gurnett, D., & Hord, C. 1993, Natur, 365, 715

Savransky, D., & Garrett, D. 2016, JATIS, 2, 011006

Seager, S., Turnbull, M., Sparks, W., et al. 2015, Proc. SPIE, 9605, 96050W

Seo, B.-J., Gordon, B., Kern, B., et al. 2016, JATIS, 2, 011019

Skemer, A. J., Marley, M. S., Hinz, P. M., et al. 2014, ApJ, 792, 17

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757

Spergel, D., Gehrels, N., Breckinridge, J., et al. 2013, arXiv:1305.5422

Stubenrauch, C. J., Rossow, W. B., Kinne, S., et al. 2013, BAMS, 94, 1031

Toon, O. B., McKay, C. P., Ackerman, T. P., & Santhanam, K. 1989, JGR, 94, 16287

Traub, W. A., Breckinridge, J., Greene, T. P., et al. 2016, JATIS, 2, 011020

Trauger, J., Moody, D., Krist, J., & Gordon, B. 2016, JATIS, 2, 011013

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, arXiv:1102.1523

von Paris, P., Hedelt, P., Selsis, F., Schreier, F., & Trautmann, T. 2013, A&A, 551, A120

Wang, J., Mawet, D., Ruane, G., et al. 2017a, Proc. SPIE, 10400, 104000Z

Wang, J., Mawet, D., Ruane, G., Hu, R., & Benneke, B. 2017b, AJ, 153, 183