Hierarchical Bayesian Atmospheric Retrieval Modeling for Population Studies of Exoplanet Atmospheres: A Case Study on the Habitable Zone

Jacob Lustig-Yaeger1,2, Kristin S. Sotzen1,3, Kevin B. Stevenson1, Rodrigo Luger2,4, Erin M. May1, L. C. Mayorga1, Kathleen Mandt1, and Noam R. Izenberg1

1 Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA
2 NASA NExSS Virtual Planetary Laboratory, Box 351580, University of Washington, Seattle, Washington 98195, USA; Jacob.Lustig-Yaeger@jhuapl.edu
3 Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
4 Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA

Received 2021 October 18; revised 2022 January 19; accepted 2022 January 26; published 2022 February 23

Abstract

With the growing number of spectroscopic observations and observational platforms capable of exoplanet atmospheric characterization, there is a growing need for analysis techniques that can distill information about a large population of exoplanets into a coherent picture of atmospheric trends expressed within the statistical sample. In this work, we develop a Hierarchical Bayesian Atmospheric Retrieval (HBAR) model to infer population-level trends in exoplanet atmospheric characteristics. We demonstrate HBAR on the case of inferring a trend in atmospheric CO2 with incident stellar flux, predicted by the presence of a functioning carbonate–silicate weathering negative feedback cycle, an assumption upon which all calculations of the habitable zone (HZ) rest. Using simulated transmission spectra and JWST-quality observations of rocky planets with H2O, CO2, and N2 bearing atmospheres, we find that the predicted trend in CO2 causes subtle differences in the spectra of order 10 ppm in the 1–5 μm range, underscoring the challenge inherent to testing this hypothesis. In the limit of highly precise data (100 stacked transits per planet), we show that our HBAR model is capable of inferring the population-level parameters that characterize the trend in CO2, and we demonstrate that the null hypothesis and other simpler trends can be rejected at high confidence. Although we find that this specific empirical test of the HZ may be prohibitively challenging in the JWST era, the HBAR framework developed in this work may find a more immediate usage for the analysis of gas giant spectra observed with JWST, Ariel, and other upcoming missions.

Unified Astronomy Thesaurus concepts: Exoplanets (498); Exoplanet atmospheres (487); Exoplanet atmospheric composition (2021); Extrasolar rocky planets (511); Habitable zone (696); Habitable planets (695); Transmission spectroscopy (2133); Bayesian statistics (1900); Hierarchical models (1925)

1. Introduction

The habitable zone (HZ) provides a tangible starting point in the search for habitable exoplanet surface environments and life beyond the solar system (Kasting et al. 1993; Kopparapu et al. 2013; Kaltenegger 2017; Meadows & Barnes 2018). Despite the HZ’s scientific lineage rooted in Earth system science (e.g., Hart 1978, 1979), understanding the persistent habitability of Earth over geologic time remains an ongoing interdisciplinary investigation (e.g., Goldblatt & Zahnle 2011; Charnay et al. 2020; Isson et al. 2020; Stüeken et al. 2020), even before the principles of Earth’s habitability are extended into the lesser understood exoplanet sample. Rather than underestimating the usefulness of the HZ, these underlying Earth-centric assumptions form the basis of a compelling hypothesis on the general nature of planetary habitability that has yet to be observationally tested using exoplanets, and may one day feed back into our understanding of Earth (Komacek et al. 2021; Shorttle et al. 2021)

This connection is well-illustrated by the carbonate–silicate weathering negative feedback cycle (Walker et al. 1981), which is thought to have helped maintain habitable surface conditions on Earth over billions of years via atmospheric CO2 buffering (Berner 2003), but which is also an assumed ingredient in HZ calculations (Kasting et al. 1993; Williams & Kasting 1997; Kopparapu et al. 2013). In the carbonate–silicate weathering cycle, atmospheric CO2 warms the climate via the greenhouse effect. An increase in volcanic outgassing produces more CO2, which further warms the climate. However, the weathering rate of continents also increases with temperature, thereby increasing the rate at which CO2 is removed from the atmosphere and ultimately subducted back into the mantle. Thus, the temperature dependence of the weathering rate provides the climate-stabilizing negative feedback that helps to maintain habitable surface temperatures against increases in CO2 and the solar luminosity over geologic timescales (Glaser et al. 2020).

The link between Earth’s long-term climate evolution and our perspective on exoplanet habitability provides a compelling opportunity to observationally test such hypotheses on the nature of planetary habitability using the population of exoplanets. Bean et al. (2017) outlined a statistical comparative planetology approach to empirically test the HZ hypothesis by recognizing that the climate model calculations for the HZ form a set of predictions that can be tested in the future using the growing sample of known likely rocky exoplanets. Specifically, if the carbonate–silicate weathering feedback mechanism operates roughly as expected by climate theory, "rocky exoplanets should exhibit a trend of increasing atmospheric CO2 from the inner edge of the HZ to the outer edge such that temperate surfacetemperatures are maintained. Lehmer et al. (2020) used a coupled climate and weathering model to investigate the dependence of this trend on practical geophysical and physiochemical differences that are likely to
exist between exoplanets. They reaffirmed the existence of such a trend in their models, but found that significant scatter may make it difficult to distinguish observationally, instead suggesting that the 2D distribution of planets in the flux-CO₂ phase space may yield a more reliable test of the HZ hypothesis. Thus, although no single exoplanet can offer a definitive test of the HZ, the entire exoplanet ensemble provides a population that, in theory, can.

However, testing for a statistical comparative planetology trend in atmospheric composition will be challenging because the trend itself is not directly observable, but must be properly identified by synthesizing the results of many individual inferences. Since exoplanet atmospheric compositions must be inferred from spectroscopic observations using retrieval models, any trends in atmospheric composition fall into the category of multilevel or hierarchical inference problems. While numerous trends in exoplanet atmospheric composition have been suggested as a means to understand exoplanet habitability (e.g., Checlair et al. 2019; Turbet et al. 2019; Bixel & Apai 2020, 2021; Checlair et al. 2021), a consistent framework to tackle the hierarchical atmospheric retrieval problem—going from spectroscopic observations to population-level atmospheric trends—has not been presented.

In this paper, we present a novel retrieval methodology that enables inferences from observations of multiple planets to be combined and synthesized to constrain population-level atmospheric characteristics, and we apply the model to an idealized population of potentially habitable exoplanets to test for the predicted carbonate–silicate weathering CO₂ trend. This is achieved using a new hierarchical Bayesian atmospheric retrieval (HBAR) modeling approach using the importance sampling formalism from Hogg et al. (2010). This use case is a natural extension of classical hierarchical Bayesian parameter estimation problems (Gelman et al. 2013), which have been highly successful in numerous exoplanet population studies (e.g., Hogg et al. 2010; Rogers 2015; Wolfgang et al. 2016) and recently applied to the atmospheric characterization of hot Jupiters using Spitzer eclipse measurements (Keating & Cowan 2022). However, such methods have yet to be implemented for exoplanet atmospheric retrievals. Typically, hierarchical models are used to properly account for and determine the underlying population-level prior distributions from which an entire population of astrophysical objects are sampled. For instance, the mass–radius relationship has been constrained for sub-Neptune-sized planets using mass and radius inferences across an ensemble of known exoplanets (Wolfgang et al. 2016). While completely novel to exoplanet atmospheric retrievals, a similar approach has been applied to Earth remote sensing aerosol retrievals by leveraging a hierarchical model with a built-in spatial dependence to capture spatial smoothness (Wang et al. 2011).

In Section 2, we describe our standard and hierarchical retrieval methods. In Section 3, we present the results of our atmospheric retrieval modeling, and then use them to infer CO₂ trends with our HBAR model and perform a population-level model comparison. In Section 4 and Section 5, we provide a discussion and the conclusions of our findings, respectively.

2. Methods

We begin in Section 2.1 by describing our nominal exoplanet atmospheric retrieval model, which shares many common traits with other retrieval codes in the literature. We then detail the hierarchical modeling approach and how it interfaces with the standard retrieval framework in Section 2.2.

2.1. Nominal Retrieval Model

We use the Spectral Mapping Atmospheric Radiative Transfer for Exoplanet Retrieval model (smarter) to solve the Bayesian inverse problem on simulated terrestrial exoplanet transmission spectra (Lustig-Yaeger 2020; Lustig-Yaeger et al. 2022, in preparation). Below, we provide a brief description of the smarter model, which is limited to the essential components for this work, but we refer the reader to Lustig-Yaeger (2020) and Lustig-Yaeger et al. (2022, in preparation) for a complete description of the smarter retrieval model and its rigorous validation using exoplanet-analog observations of Earth’s infrared transmission spectrum.

2.1.1. Forward Model

smarter relies on the Spectral Mapping Atmospheric Radiative Transfer model (smart) as the core of the forward model used to simulate line-by-line transmission spectra for transiting exoplanets (Meadows & Crisp 1996; Crisp 1997; Misra et al. 2014) using the ray tracing formalism described in Robinson (2017). In turn, smart leverages the DIScrete Ordinate Radiative Transfer (DISORT; Stamnes et al. 2017) model to solve the radiative transfer equation. The Line-By-Line Absorption Coefficient code (lblabc; developed by D. Crisp; Meadows & Crisp 1996) is used to calculate molecular vibrational-rotational absorption coefficients for input into smart radiative transfer calculations. lblabc combines information about the atmospheric state with HITRAN line-parameter and isotope information from the HITRAN2016 line list (Gordon et al. 2017) to calculate gas absorption coefficients as a function of pressure, temperature, and wavenumber. Collisionally induced absorption data are used for CO₂ – CO (Moore 1972; Kasting et al. 1984; Gruszka & Borysow 1997; Baranov et al. 2004; Wordsworth et al. 2010; Lee et al. 2016) and N₂ – N₂ (Lafferty et al. 1996; Schwieterman et al. 2015b).

For simplicity in this study, we consider one-dimensional atmospheres composed of N₂, CO₂, and H₂O with isothermal temperature–pressure profiles and evenly mixed gas abundances. While plainly limited, this combination of gases is consistent with the climate modeling work that underpins the HZ (e.g., Kasting et al. 1993; Kopparapu et al. 2013). We allow the (log₁₀) volume mixing ratios (VMRs) of CO₂ and H₂O to freely vary within the forward model, but set the N₂ VMR to the residual VMR such that the VMRs of all three gases sum to unity, as in numerous retrieval studies (e.g., Feng et al. 2018; Krissansen-Totton et al. 2018; Barstow et al. 2020). In addition to the (log₁₀) VMRs of CO₂ and H₂O, we fit for the isothermal temperature \( T_0 \) (in Kelvin), the solid-body surface reference planet radius \( R_0 \) (in \( R_\oplus \)), and the surface reference pressure \( P_0 \) (in log₁₀ Pascals). Although we do not formally include clouds in this study, the reference radius and pressure can be used to account for an opaque gray cloud top. In total, we use a five-parameter state vector, \( \omega \), for our forward model, subject to the following uninformative priors on \( \omega \):

\[
\begin{align*}
\ell_{H_2O} &\sim \mathcal{U}((-12, 0) \log_{10}(\text{VMR}) \\
\ell_{CO_2} &\sim \mathcal{U}((-12, 0) \log_{10}(\text{VMR}) \\
T_0 &\sim \mathcal{U}(50, 500) \text{ K} \\
R_0 &\sim \mathcal{U}(0.95, 1.05) \; R_\oplus \\
P_0 &\sim \mathcal{U}(2, 6) \log_{10}(\text{Pa})
\end{align*}
\]

(1)
where $\mathcal{U}$ (lower, upper) denotes a uniform distribution with finite probability between the lower and upper bounds. We use the function $g(\omega)$ to denote the forward model transformation of parameters $\omega$ into wavelength dependent spectroscopic units that can be directly compared to the data.

### 2.1.2. Inverse Model

We use the dynesty nested sampling code (Skilling 2004; Speagle 2020) to solve the Bayesian inverse problem for the posterior probability distribution function (PDF) of our forward model parameters given transmission spectrum observations. A standard $\chi^2$ log-likelihood function is used to calculate the probability of the transmission spectrum data ($\Delta F = (R_f/R_s)^2$) given the model parameters,

$$\ln L = -\frac{1}{2} \sum_{j=1}^{M} \left( \frac{\Delta F_j - g(\omega)}{\sigma_j} \right)^2,$$

where $\sigma_j$ is uncertainty on the spectrum for the $j$th observed wavelength. Adding the log-likelihood to the logarithm of the remaining prior volume to the total evidence to be achieved when the estimated contribution of the wavelength. Adding the log-likelihood to the logarithm of the prior PDF

of our forward model parameters gives the model parameters given transmission spectrum observations. A standard $\chi^2$ log-likelihood function is used to calculate the probability of the transmission spectrum data ($\Delta F = (R_f/R_s)^2$) given the model parameters,

$$\ln L = -\frac{1}{2} \sum_{j=1}^{M} \left( \frac{\Delta F_j - g(\omega)}{\sigma_j} \right)^2,$$

where $\sigma_j$ is uncertainty on the spectrum for the $j$th observed wavelength. Adding the log-likelihood to the logarithm of the remaining prior volume to the total evidence to be achieved when the estimated contribution of the wavelength yields $K$ equally weighted samples from the posterior distribution, where $K \approx 16,000$ for our experimental setup, as we will see in Section 3.2.

### 2.2. Hierarchical Modeling Approach

We employ the importance sampling hierarchical model described in Hogg et al. (2010) originally presented for inferring the eccentricity distribution of exoplanets. This model has numerous advantages over a traditional fully coupled hierarchical approach. First, all atmospheric retrievals can be precomputed independently and potentially in parallel, using traditional methods and posterior sampling approaches. This assumes that there are no likelihood covariances between parameters from different planets, but benefits computationally from not being required to sample the corresponding high-dimensional spaces.

Following closely with the derivations presented in Hogg et al. (2010), for any exoplanetary spectrum $n$, there are $\omega_n$ parameters that we attempt to infer,

$$\omega_n \equiv \{ f_{\mathrm{H_2} \mathrm{O}}, f_{\mathrm{CO}_2} \mathrm{, } T_n, R_n, P_n \},$$

as defined in Section 2.1.

Consider that we have $N$ exoplanets $n$ ($1 \leq n \leq N$), each of which has $M_n$ transmission spectrum measurements $\Delta F_{nj}$ (or wavelength resolution elements). For every exoplanet $n$, the set of spectroscopic measurements

$$D_n \equiv \{ \Delta F_{nj} \}_{j=1}^{M_n}$$

is modeled using a radiative transfer forward model. This is given by

$$\Delta F_{nj} = \left( \frac{R_j}{R_n} \right)^2 = g_n(\omega_n) + N(0, \sigma_{nj}^2),$$

where the function $g_n(\omega_n)$ is the spectroscopic forward model described previously in Section 2.1.1, which is parameterized in terms of the five aforementioned dimensions of $\omega_n$, and has an additive noise component drawn from a normal distribution ($\mathcal{N}$) with variance $\sigma_{nj}^2$ (the uncertainty of the $j$th observed wavelength of the $n$th star-planet system). The model of all $N$ planets has $5 \times N$ continuous parameters in the larger list $\{ \omega_n \}_{n=1}^N$. We note that it is the high-dimensional $5 \times N$ parameter space that makes a fully coupled hierarchical model (e.g., using PyMC3) computationally inefficient, as it must simultaneously infer all $5 \times N$ parameters.

The likelihood $L_n$ for the five parameters $\omega_n$ for the $n$th exoplanet spectrum is the probability of the data $D_n$ for planet $n$ given the parameters $\omega_n$:

$$L_n \equiv \mathcal{P}(D_n | \omega_n).$$

This is the standard Bayesian likelihood function previously defined in Section 2.1.2. Now for each planet $n$, suppose that we have inferred (using the inverse model described in Section 2.1.2) or been provided with a $K$-element sample from a posterior PDF created from the likelihood and an uninformative prior PDF $\mathcal{P}_0(\omega_n)$:

$$\mathcal{P}(\omega_n | D_n) = \frac{\mathcal{P}(D_n | \omega_n) \mathcal{P}_0(\omega_n)}{Z_n},$$

where $Z_n$ is the marginal likelihood or evidence, and is simply a normalization constant. We expect that the prior PDF $\mathcal{P}_0(\omega_n)$ is uninformative. For every exoplanet $n$, this posterior sampling is a collection of $K$ equally weighted samples $k$, each itself a set of five parameters $\omega_{nk}$. The joint likelihood $L$ of all parameters for all exoplanets $n$ in the dataset is simply the product of the individual likelihoods:

$$L \equiv \mathcal{P}(D_n \equiv \{ \omega_n \})_{n=1}^N \equiv \prod_{n=1}^{N} L_n.$$

As stated in Hogg et al. (2010), this makes the assumption that there are no likelihood covariances between the parameters of different exoplanets $n$.

Now we want to reframe the problem slightly and instead consider the likelihood $L_n$ for the set of (hyper)parameters $\alpha$ that are used to define an updated prior probability on the CO$_2$ abundance $\mathcal{P}_0(f_{\mathrm{CO}_2})$. It is this updated prior that will be used to encode and characterize population-level trends in atmospheric parameters, and once known, is a better choice of prior on CO$_2$ than our original uninformative choice. The probability of the entire data ensemble given the population-level parameters $\alpha$ is then given by:

$$L_\alpha \equiv \mathcal{P}(D_n \equiv \{ \omega_n \})_{n=1}^N | \alpha).$$

This joint likelihood can be expressed as the product of $N$ marginalization integrals over parameters $\omega_n$,

$$L_\alpha = \prod_{n=1}^{N} \int \mathcal{P}(D_n, \omega_n | \alpha) d\omega_n \equiv \prod_{n=1}^{N} \int \mathcal{P}(D_n | \omega_n, \alpha) \mathcal{P}(\omega_n | \alpha) d\omega_n,$$

where $\mathcal{P}(D_n | \omega_n, \alpha)$ is the spectroscopic forward model described previously in Section 2.1.1, which is parameterized in terms of the five aforementioned dimensions of $\omega_n$, and has an additive noise component drawn from a normal distribution ($\mathcal{N}$) with variance $\sigma_{nj}^2$ (the uncertainty of the $j$th observed wavelength of the $n$th star-planet system). The model of all $N$ planets has $5 \times N$ continuous parameters in the larger list $\{ \omega_n \}_{n=1}^N$. We note that it is the high-dimensional $5 \times N$ parameter space that makes a fully coupled hierarchical model (e.g., using PyMC3) computationally inefficient, as it must simultaneously infer all $5 \times N$ parameters.

The likelihood $L_n$ for the five parameters $\omega_n$ for the $n$th exoplanet spectrum is the probability of the data $D_n$ for planet $n$ given the parameters $\omega_n$:

$$L_n \equiv \mathcal{P}(D_n | \omega_n).$$

This is the standard Bayesian likelihood function previously defined in Section 2.1.2. Now for each planet $n$, suppose that we have inferred (using the inverse model described in Section 2.1.2) or been provided with a $K$-element sample from a posterior PDF created from the likelihood and an uninformative prior PDF $\mathcal{P}_0(\omega_n)$:

$$\mathcal{P}(\omega_n | D_n) = \frac{\mathcal{P}(D_n | \omega_n) \mathcal{P}_0(\omega_n)}{Z_n},$$

where $Z_n$ is the marginal likelihood or evidence, and is simply a normalization constant. We expect that the prior PDF $\mathcal{P}_0(\omega_n)$ is uninformative. For every exoplanet $n$, this posterior sampling is a collection of $K$ equally weighted samples $k$, each itself a set of five parameters $\omega_{nk}$. The joint likelihood $L$ of all parameters for all exoplanets $n$ in the data set is simply the product of the individual likelihoods:

$$L \equiv \mathcal{P}(D_n \equiv \{ \omega_n \})_{n=1}^N \equiv \prod_{n=1}^{N} L_n.$$

As stated in Hogg et al. (2010), this makes the assumption that there are no likelihood covariances between the parameters of different exoplanets $n$.

Now we want to reframe the problem slightly and instead consider the likelihood $L_n$ for the set of (hyper)parameters $\alpha$ that are used to define an updated prior probability on the CO$_2$ abundance $\mathcal{P}_0(f_{\mathrm{CO}_2})$. It is this updated prior that will be used to encode and characterize population-level trends in atmospheric parameters, and once known, is a better choice of prior on CO$_2$ than our original uninformative choice. The probability of the entire data ensemble given the population-level parameters $\alpha$ is then given by:

$$L_\alpha \equiv \mathcal{P}(D_n \equiv \{ \omega_n \})_{n=1}^N | \alpha).$$

This joint likelihood can be expressed as the product of $N$ marginalization integrals over parameters $\omega_n$,
which can be factored further by assuming that the data $D_n$ depend on $\alpha$ only through $\omega_n$ (i.e., $P(D_n|\omega_n, \alpha) = P(D_n|\omega_n)$), such that

$$L_\alpha = \prod_{n=1}^N \int P(D_n|\omega_n)P(\omega_n|\alpha)d\omega_n. \quad (13)$$

We recognize the first term in the integrand as the individual likelihood for the $n$th planet spectrum, but the second term—the probability of $\omega_n$ given $\alpha$—is critical and allows us to reweight the integral using our new hierarchical prior,

$$P(\omega_n|\alpha) = \frac{P(f_{\text{CO}_2}|\alpha)P(\omega_n)}{P(f_{\text{CO}_2})}. \quad (14)$$

Equation (14) divides out the contribution from CO2 to the original uninformative prior and multiplies through by the new CO2 prior that depends on the hyperparameters $\alpha$. Substituting Equation (14) into Equation (13) and recognizing that the product of the individual likelihood times the original uninformative prior is the posterior via Equation (7), we arrive at the following $N$ multidimensional integrals:

$$L_\alpha = \prod_{n=1}^N \int P(\omega_n|D_n)\frac{P(f_{\text{CO}_2}|\alpha)}{P(f_{\text{CO}_2})}d\omega_n. \quad (15)$$

Note that we have dropped the dependence of Equation (15) on the marginal likelihood $Z_n$ because typical posterior inference applications only evaluate Equation (7) up to the unknown normalization constant.

While Equation (15) looks computationally exhausting, the fact that we have already obtained posterior samples simplifies the integral substantially. As articulated in Hogg et al. (2010), since all probability integrals can be approximated as sums over samples, we can employ the posterior sampling approximation to obtain

$$L_\alpha \approx \prod_{n=1}^N \frac{1}{K} \sum_{k=1}^K \frac{P(f_{\text{CO}_2,k})}{P(f_{\text{CO}_2})}. \quad (16)$$

where $k$ runs over all posterior samples $K$, and the sum simply contains the ratio of the new prior PDF that we want to infer $P(f_{\text{CO}_2}|\alpha)$ to the uninformative prior PDF that was used in the original retrieval inference. Although the individual posteriors $P(\omega_n|D_n)$ do not explicitly appear in Equation (16), they are implicitly contained within the distribution of $K$ samples. With the likelihood $L_\alpha$ as defined in Equation (16), it is straightforward to infer posteriors on the population-level parameters $\alpha$ using Bayes’ Theorem,

$$P(\alpha|D_n) \propto L_\alpha P(\alpha), \quad (17)$$

where $P(\alpha)$ is the (hyper) prior PDF for the hyperparameters $\alpha$.

Within this section, we have kept the importance sampling derivations as agnostic as possible to the specifics of the population-level trend(s) under consideration, to ensure that the methods can be readily adapted to other problems. Critically, we have not yet specified the form of the CO2 trend, the hyperparameters that define $\alpha$, or their respective hyperpriors.

5 This is a common assumption in hierarchical Bayesian inference, since population-level hyperparameters are typically used to constrain the priors on the physical parameters at the individual level, rather than to directly modify the observables.

We refer the reader to Section 3 (in particular, Equation (18), Equation (19), and Equation (20)) for our specific implementation and the subsequent results.

In the original Hogg et al. (2010) formulation of importance sampling, it was assumed that the original exoplanet data was not in hand, but that the posteriors had been obtained from a colleague or another research group for further analysis. While this is certainly a circumstance that may motivate the use of importance sampling for atmospheric retrievals, we found that it was crucial, perhaps necessary, to perform the hierarchical analysis in a subsequent step following the completion of a uniform set of retrievals, due to the excessive computational expense of simultaneously inferring the population parameters $\alpha$ along with all of the $5 \times N$ individual system atmospheric parameters $\omega_n$. This intractability stems from the inherent computational expense of retrieval codes, which solve the radiative transfer equation at each step in the spectral inference. However, retrieval codes with exceptionally fast forward models may prove important for exploring the cost–benefit analysis of HBAR methods that simultaneously infer individual and population parameters. This may be an opportunity for machine-learning-augmented retrievals (e.g., Zingales & Waldmann 2018; Hayes et al. 2020; Himes et al. 2020; Nixon & Madhusudhan 2020).

3. Results

First, in Section 3.1, we present a uniform set of spectral models for an idealized population of terrestrial exoplanets that by design exhibit the silicate weathering CO2 trend of interest. Second, in Section 3.2, we use our new HBAR model to infer the atmospheric CO2 trend across the population of synthetic exoplanets. Third, in Section 3.3, we conduct a population-level model comparison to determine the robustness of the inferred trend relative to the null hypothesis and other functional forms for the silicate weathering relation.

3.1. Spectral Models

We generate a set of transmission spectra that will allow us to empirically test for the existence of the HZ as described in Bean et al. (2017). To limit the number of confounding factors in this study, we assume that the set of $N$ exoplanets with observed spectra are identical to one another except for their stellar irradiation and the quantities of CO2 and N2 in their atmospheres, which follows the silicate weathering feedback trend that we impose. We assume the planets possess 1 bar Earth-like atmospheres composed only of N2, CO2, and H2O. We use globally averaged Earth vertical thermal and H2O profiles (from Robinson et al. 2010, 2011; Schwieterman et al. 2015a) to satisfy the assumption that each planet is habitable, but we neglect changes in these profiles that would be expected from self-consistent climate modeling across the HZ, to limit our focus to observables due to CO2. Furthermore, we assume that all simulated planets are Earth-sized (1 $M_e$, 1 $R_e$) and orbit TRAPPIST-1 with the same transit duration as TRAPPIST-1e (Gillon et al. 2017; Agol et al. 2021), which simply provides a tangible point of comparison to judge the plausibility of such an analysis in the context of current exoplanet targets and observing capabilities.

To simplify the underlying model for our population trend, we fit an analytic function to the predicted CO2 VMRs calculated by Bean et al. (2017) using a 1D radiative-convective climate
We used the following “Gaussian-like” functional form

\[
f_{\text{CO}_2}(\mu, \sigma, \delta) = \frac{1}{\delta \sqrt{2\pi}} \exp \left[ -0.5 \left( \frac{S_0 - \mu}{\sigma} \right)^4 \right].
\]  

(18)

where \( S_0 \) is the stellar irradiation incident on the planet relative to Earth and \( \mu, \sigma, \) and \( \delta \) are free parameters that we determine to be 0.04727, 0.5372, and 0.4376, respectively, by minimizing the squared residuals. We assume that the remainder of the atmospheric volume is filled with \( N_2 \), and then we calculate the mean molecular weight of the atmosphere self-consistently.

Figure 1 shows our resulting transmission spectrum models at 1 cm\(^{-1}\) wavenumber resolution (left panel) that correspond to the assumed trend in \( \text{CO}_2 \) with stellar irradiation (right panel) due to the carbonate–silicate weathering feedback mechanism for \( N = 20 \) theoretical exoplanets. By design, the spectra exhibit differences due solely to the VMR of \( \text{CO}_2 \) (and implicitly \( N_2 \)), which are small relative to the total transit depth. Two competing effects shape the observable characteristics of the spectra shown in Figure 1: the \( \text{CO}_2 \) optical depth and the atmospheric mean molecular weight. As the \( \text{CO}_2 \) abundance increases, the \( \text{CO}_2 \) optical depth increases, and the weak \( \text{CO}_2 \) bands, primarily seen between 1 and 2 \( \mu \text{m} \), increase in absorption strength. The opposite is seen for the saturated \( \text{CO}_2 \) bands at 2.7 and 4.3 \( \mu \text{m} \). The increase in \( \text{CO}_2 \) causes the saturated bands to decrease in absorption strength as the mean molecular weight of the atmosphere increases from \( N_2 \)-dominated (28 g mol\(^{-1}\)) to \( \text{CO}_2 \)-dominated (44 g mol\(^{-1}\)), and the atmospheric scale height decreases correspondingly. In general, these subtle spectral differences must be sufficiently resolved in each observed spectrum for the population-level model to infer a meaningful trend.

We used the PandExo JWST noise model (Batalha et al. 2017, 2018) to simulate synthetic transmission spectrum observations using the Near-Infrared Spectrograph (NIRSpec) Prism instrument (Bagnasco et al. 2007; Ferruit et al. 2014). We used the same PandExo simulation setup as Lustig-Yaeger et al. (2019) assuming the partial saturation strategy for the NIRSpec Prism (Batalha et al. 2018) and no assumed noise floor. Figure 2 shows the precision of the Prism spectra for TRAPPIST-1e using 10 and 100 stacked transits, compared against the standard deviation among the spectra shown in Figure 1. Based on the fact that the 1\( \sigma \) spectral uncertainties for 100 transits are of similar magnitude to the deviations caused by \( \text{CO}_2 \), we conclude that approximately 100 observed transits may be required, for each planet, to obtain spectra with high enough precision to clearly resolve the deviations in \( \text{CO}_2 \) mixing ratios across HZ instellations.

Figure 1. Transmission spectrum models (left) for a sample of 20 hypothetical rocky exoplanets generated following the trend in \( \text{CO}_2 \) with stellar irradiation predicated on the assumption of a functioning carbonate–silicate weathering feedback mechanism (right). Planets near the outer edge of the HZ at lower stellar irradiation are predicted to possess atmospheres with higher \( \text{CO}_2 \) abundances than those near the inner edge of the HZ, and such variations in \( \text{CO}_2 \) manifest in observable features in the transmission spectrum.

Figure 2. Standard deviation among the spectra shown in Figure 1 with varying \( \text{CO}_2 \) (black) compared with the 1\( \sigma \) observational uncertainties for 10 (magenta) and 100 (orange) stacked transits of TRAPPIST-1e with JWST’s NIRSpec Prism. The spectroscopic differences caused by the silicate weathering \( \text{CO}_2 \) trend are less than 10 ppm for Earth-sized exoplanets transiting TRAPPIST-1, indicating that approximately 100 stacked transits may be necessary to clearly resolve the deviations in \( \text{CO}_2 \) mixing ratios across HZ instellations.
3.2. An Empirical Test of the Habitable Zone

We performed a uniform set of retrievals on the 20 spectra shown in Figure 1 that evenly span the HZ range of stellar irradiation with spectral resolution and noise appropriate for 100 transits with JWST/NIRSpec Prism. Figure 3 shows a representative corner plot of the 1D and 2D marginalized posterior distributions for the inferred physical planetary parameters for the case with \( \sim 83\% \) atmospheric CO\(_2\)—the fourth largest CO\(_2\) abundance in the sample. The covariance between the isothermal temperature and CO\(_2\) abundance shows a degeneracy due to the dependence of the atmospheric scale height on the temperature and mean molecular weight. Similarly, the reference radius and pressure show an expected degeneracy that represents the set of radii and pressures that maintain the transmission spectrum continuum near the observed value. The retrieved H\(_2\)O abundance is consistent with stratospheric values, and notably for the purpose of this investigation, the CO\(_2\) abundance is constrained to within ±0.5 dex. The upper right inset shows the median retrieved spectrum bounded by 1\( \sigma \) and 3\( \sigma \) credible intervals derived from posterior samples. This particular retrieval result was for the case with \( \sim 83\% \) CO\(_2\) and is representative of similar results obtained for the other planets in our synthetic sample.

Figure 3. Corner plot showing the 1D and 2D marginalized posterior distributions (histograms and contours) for the five planetary parameters retrieved from fitting the transmission spectrum using the smarter model. The subplots along the diagonal show the input atmospheric profiles as a function of pressure that were used to generate the synthetic spectrum for comparison against the vertically homogeneous (isothermal and evenly mixed) atmospheric models that were retrieved. The upper right inset shows the median retrieved spectrum bounded by 1\( \sigma \) and 3\( \sigma \) credible intervals derived from posterior samples. This particular retrieval result was for the case with \( \sim 83\% \) CO\(_2\) and is representative of similar results obtained for the other planets in our synthetic sample.
Although the retrieval results shown in Figure 3 are only for one representative case from our sample, the other 19 retrievals show similar results with expected differences caused by the different underlying CO2 abundance and the propagation of random Gaussian scatter in the spectrum through the inference procedure. On average, each retrieval with dynesty yielded $K \approx 16,000$ equally weighted samples from the respective posteriors. Next, the ensemble of posteriors obtained from the individual planets will be used to infer population-level parameters, in an attempt to retrieve the silicate weathering feedback trend that we injected into the sample.

With the posteriors from our uniform retrieval analysis in hand, we now wish to infer population-level parameters for the CO2 versus stellar irradiation trend by running MCMC on the HBAR importance sampling model. The HBAR likelihood function is given by Equation (16), where the original prior $\mathcal{P}_0(\hat{f}_{CO2,i})$ on the log CO2 abundance is uninformative $\mathcal{U}(-12, 0)$ and the updated prior $\mathcal{P}_\alpha(\hat{f}_{CO2})$ is calculated from the analytic relationship provided in Equation (18). Specifically, the updated prior is a function of hyperparameters $\alpha$ and is taken to be normally distributed,

$$\mathcal{P}_\alpha(\hat{f}_{CO2}) = \mathcal{N}(\hat{f}_{CO2} - f_{CO2}, \alpha, \delta, \sigma_2),$$  

where $\sigma_2$ is the standard deviation of the Gaussian distribution that lends high probability to values of the CO2 population trend that lie closest to the original CO2 posterior samples. Thus, for our HBAR importance sampling model, we have the free hyperparameters $\alpha = [\mu, \sigma_1, \delta, \sigma_2]$ that we seek to infer, subject to the following uninformative hyperpriors on $\alpha$:

$$\mathcal{P}(\alpha) = \begin{cases} 
\mu & \sim \mathcal{U}(-2, 2) \\
\sigma_1 & \sim \mathcal{U}(0, 2) \\
\delta & \sim \mathcal{U}(0, 2) \\
\sigma_2 & \sim \mathcal{N}_1/\sqrt{2}(0, 1),
\end{cases}$$

where $\mathcal{N}_1/\sqrt{2}(0, 1)$ refers to a half-normal distribution with a mean of 0 and a standard deviation of 1. At this point, we are ready to evaluate our HBAR model and infer the hyperparameters $\alpha$. To recap, we now have a specific population-level model for the CO2 trend (Equation (18)) that is used to define a new population-level prior on the CO2 abundance (Equation (19)). This allows the likelihood $\mathcal{L}_\alpha$ to be obtained by evaluating Equation (16). Finally, the uninformative hyperprior (Equation (20)) can be multiplied by the likelihood, as in Equation (17), to infer the desired $\alpha$ posterior PDF.

We used MCMC with emcee (Foreman-Mackey et al. 2013) to infer posterior samples of the hyperparameters $\alpha$ using 20 walkers. We ran the chain until it reached a length of approximately $50 \times$ the integrated autocorrelation time, as suggested by the emcee code documentation.6

Figure 4 shows the MCMC results from our HBAR importance sampling model. The lower left set of panels in Figure 4 show the 1D and 2D marginalized posteriors for the population-level hyperparameters in our HBAR model. Relative to their uninformative hyperpriors, the hyperparameters are well-constrained by the inference. The median retrieved trend in CO2 with stellar irradiation is calculated from the posterior samples and is shown in the upper right panel of Figure 4, bounded by the 1$\sigma$ and 3$\sigma$ credible intervals. The upper right panel also shows the retrieved 1$\sigma$ CO2 constraints for all 20 independent spectrum retrievals plotted as a function of stellar irradiation. It may be conceptually useful to imagine that we have directly fit the purple population model to the black error bars in CO2 abundance, while in fact we have actually taken the full set of multidimensional posteriors into consideration in our numerical evaluation of Equation (16) and Equation (17). The true underlying CO2 trend is also shown for reference, where the characteristic decline in CO2 abundance with stellar irradiation is well-resolved by the population-level inference.

3.3. Model Comparison

One of the benefits of performing an importance sampling HBAR meta-analysis is that multiple different hypothesized population-level atmospheric trends can be investigated and compared without the need to rerun the computationally expensive retrieval models. We now compare three simpler population models to our previously obtained result, to demonstrate how such models can be discriminated. These population-level models for the CO2 versus stellar irradiation relation include a linear trend, a log-linear trend, and a flat non-trend representing the null hypothesis.

Figure 5 compares four different best-fitting models to the population-level CO2 trend. The Bayesian Information Criterion (BIC) is calculated for each model using the following relation:

$$\text{BIC} = \kappa \ln N - 2 \ln \hat{\mathcal{L}}_\alpha,$$

where $\kappa$ is the number of free population-level parameters estimated by the model (i.e., the number of dimensions in $\alpha$), $N$ is taken to be the number of individual planet spectra in the population, and $\ln \hat{\mathcal{L}}_\alpha$ is the maximum of the log-likelihood function obtained through optimization. When selecting between multiple models, the model with the lowest BIC is taken to be preferred. We subtract the preferred model BIC from every other model to obtain $\Delta$BIC values for comparison. The $\Delta$BIC values in Figure 5 indicate that the “Gaussian-like” model is preferred and there is strong evidence against all of the models with higher BICs. This is the expected result because we generated the synthetic planetary models using the Gaussian-like population trend, but this serves as a useful demonstration of how population-level trends can be compared using an importance sampling HBAR framework.

4. Discussion

We conducted an idealized simulated search for a trend in CO2 abundance with stellar irradiation predicted by the carbonate–silicate weathering feedback for exoplanets in the HZ. Our approach used a novel hierarchical Bayesian model—a first of its kind for exoplanet atmospheric retrievals—to infer population trends in atmospheric characteristics that may prove useful well beyond the scope of this work. We elaborate on our scientific and methodological findings in the following subsections.

4.1. Practical Challenges to an Empirical Test of Carbonate–Silicate Weathering

We found that the CO2 trend predicted by carbonate–silicate weathering will be challenging to infer, potentially limiting the observational feasibility of this statistical comparative planetology test of the HZ. As presented in Bean et al. (2017), the
use of a large statistical sample of exoplanets is enabled by making relatively low-precision spectroscopic observations of each planet. However, as we have shown, the changes to the transmission spectrum of an Earth-like planet (which this investigation must detect) caused by changes in CO₂ abundance (which are consistent with the predicted trend) are quite small at <10 ppm for TRAPPIST-1e-like planets. As a result, to infer the silicate weathering trend at high confidence, we found that precise transmission spectra were required that corresponded to 100 stacked transit observations with JWST of each planet in our idealized 20 planet sample, despite the optimistic assumption that each planet was TRAPPIST-1e-like (while only one such planet is known to exist). Acquiring such a sample of high-precision terrestrial exoplanet transmission spectra is not feasible in JWST’s nominal mission lifetime, nor would it be made appreciably easier using a successor to
direct-imaging mission that can obtain spectra of Earth-like exoplanets around Sun-like stars (as recommended by the National Academies of Sciences, Engineering, and Medicine 2021), such as the Large UV/Optical/IR Surveyor (LUVOIR; LUVOIR Mission Concept Study Team 2019) or the Habitable Exoplanet Observatory (HabEx; HabEx Study Team 2020) with access to the 1.6 μm and 2 μm CO2 bands, may offer more leverage for precise CO2 abundance retrievals. However, this will need to be demonstrated in a future study, because these weak CO2 bands were omitted from the seminal retrieval work of Feng et al. (2018), due to the relative insignificance of CO2 in the Earth’s visible and NIR spectrum.

Nature is likely to produce more complicated atmospheric trends than what we have investigated here. This may hold particularly true for habitable exoplanets where the presence of stable surface liquid water is appreciated to be dependent on a complex interplay of stellar, planetary, and planetary-system-wide factors that may indeed produce an elusive population trend that spans many dimensions (Meadows & Barnes 2018). To this end, recent work by Lehmer et al. (2020) demonstrated that the carbonate–silicate weathering feedback trend in CO2 with incident flux may be log-linear in form with significant scatter due to individual planet considerations, such as land area for weathering and CO2 outgassing fluxes. The log-linear trend differs from the nonlinear trend from Bean et al. (2017) considered in this work because the Lehmer et al. (2020) model included temperature and CO2 feedbacks that cause the surface temperature to decline with semimajor axis (Kadoya & Tajika 2014), rather than remain fixed at 289 K throughout the HZ (Bean et al. 2017). The exact nature of the predicted CO2–flux trend in the HZ is unlikely to change our results, due to the relative consistency of the two similar hypotheses compared to our posterior constraints on CO2. Moreover, these predictions are all likely to be incorrect at some level, due to their exclusive reliance on geophysical evidence from Earth. This only further motivates the need for methods that allow us to update our understanding of comparative planetology trends using exoplanet data. Future work could leverage the HBAR framework presented here to investigate the 2D population density trend suggested by Lehmer et al. (2020) and potentially incorporate a third dimension for the surface temperature to better capture the climatic feedbacks expected throughout the HZ (Kadoya & Tajika 2014; Lehmer et al. 2020).

Similarly, Seales & Lenardic (2021) used coupled geophysical models to study the temporal onset of habitability and found that variations in tectonic efficiency from one planet to another may produce a predictable distribution in the CO2 abundance for an ensemble of planets with the same absolute age. Thus, it may be possible to expand the population trends studied here into the system age dimension to relate an inferred distribution in CO2 back to predictions from geophysical models.

4.2. The Future of HBAR

We have taken a first step toward developing a hierarchical Bayesian atmospheric retrieval model for tracking population trends in exoplanet atmospheres through the complicated, nonlinear, and degenerate problem of fitting exoplanet spectra. Standard atmospheric retrievals are well-known to be computationally expensive due to the requirement of a radiative transfer forward model to fit spectroscopic observations. By using the relatively simple importance sampling method for
hierarchical Bayesian modeling from Hogg et al. (2010), we effectively avoid the computationally taxing need to perform retrievals on each planet’s spectrum simultaneously, as would be required for a standard hierarchical model. Instead, importance sampling allows for population-level inferences in a straightforward meta-analysis of the posterior samples obtained from a uniform set of standard atmospheric retrieval results. This effectively decouples the computationally expensive retrieval modeling from the hierarchical modeling, such that the hierarchical problem can be readily solved once a uniform set of retrieval results (posteriors) are in hand. Thus, intrigued readers may find that they already have all of the necessary ingredients to characterize population-level trends within their existing retrieval results.

Characterizing population-level trends in exoplanet atmospheres has use cases well beyond the HZ and offers a critical capability for advancing comparative planetology with current, upcoming, and future telescopes. For example, population studies of extrasolar gas giants are already underway. Seminal work by Sing et al. (2016) analyzed the spectra of 10 hot Jupiters observed with HST and found that the planets exhibited a continuum from clear to cloudy atmospheres, which may suggest that clouds and hazes, rather than water depletion during formation, are the cause of weaker-than-expected H$_2$O absorption features. However, subsequent uniform retrieval analyses by Barstow et al. (2017) and Pinhas et al. (2019) complicate this picture, as their retrieved H$_2$O abundances suggest subson oxygen and/or supersolar C/O ratios with no clear correlations identified. Additionally, Tsiaras et al. (2018) conducted a population study of 30 gaseous exoplanets and found that about half of the sample had detectable atmospheres via H$_2$O absorption features. Future work on this hot Jupiter sample may benefit from the HBAR model described here to infer and compare population trends for different proposed formation and evolutionary pathways. Moving toward smaller planets, Changeat et al. (2020) conducted a uniform retrieval analysis to demonstrate that the ESA-Ariel mission (Tinetti et al. 2016) will be sensitive to trends between the atmospheric chemistry and planetary parameters for a population consisting of mostly sub-Neptune and Neptune-sized planets (Edwards et al. 2019). As more telescopes dedicated to exoplanet atmospheric characterization come online, and as the number of observed exoplanet spectra grows, the use of HBAR modeling may become crucial for comparative planetology.

5. Conclusion

We implemented a first-of-a-kind hierarchical Bayesian atmospheric retrieval model to characterize population-level trends in exoplanet atmospheres. We argue that hierarchical Bayesian models are well-suited for this task due to the sophisticated inference methods (retrievals) required to transform the observed spectra of exoplanets into meaningful atmospheric characteristics. In particular, the HBAR model that we implemented using importance sampling offers a computationally tractable approach for performing such multilevel inferences because it requires only the posteriors from a uniform set of traditional retrievals, which can all be performed independently. Marginalizing over the full multidimensional posteriors allows the importance sampling HBAR method to propagate complicated parameter covariances through to the population-level hyperparameters, conserving information that may be lost when analyzing retrieved atmospheric trends using only 1D marginalized posteriors or traditional statistical moments. While this study by no mean represents the end-all-be-all of HBAR modeling, we have taken the first few steps toward a computationally tractable HBAR model that will benefit over time from further application and refinement by the exoplanet community.

We tested the importance sampling HBAR framework on an empirical probe of the HZ using simulated transmission spectra of rocky planets with an injected trend in CO$_2$ abundance with stellar irradiation that is consistent with predictions for a functioning carbonate–silicate weathering negative feedback cycle. We demonstrated that the HBAR method can be used to (1) accurately constrain population-level parameters that characterize the silicate weathering trend and (2) discriminate between multiple different hypothetical population trends. However, we found that such precise spectroscopic measurements would be required to sense the CO$_2$ trend in terrestrial exoplanet atmospheres that inferring this particular statistical comparative planetology trend may be infeasible using upcoming missions with transmission spectroscopy capabilities. Nonetheless, the use of the HBAR methods presented here may prove to be an important ingredient for future comparative planetology studies as new theories of planetary atmospheric formation, evolution, and habitability are forged in the crucible of exoplanet demographics.

We would like to thank D. Foreman-Mackey for illuminating numerous dense statistical concepts, C. Tinsman for supporting our use of APL’s computing resources, and the anonymous reviewer for helping us improve the quality of this manuscript. This work was funded by internal research and development funding from the Johns Hopkins Applied Physics Laboratory. Software: Astropy (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018), corner (Foreman-Mackey 2016), Dynesty (Speagle 2020), encee (Foreman-Mackey et al. 2013), LBLABC (Meadows & Crisp 1996), Matplotlib (Hunter 2007), NumPy (van der Walt et al. 2011), SciPy (Virtanen et al. 2020), SMART (Meadows & Crisp 1996), Pandas ( McKinney 2010), PandExo (Pontoppidan et al. 2016), pysynphot (Batalha et al. 2017, 2018), and astropy (Price-Whelan et al. 2013).

ORCID iDs

Jacob Lustig-Yaeger @ https://orcid.org/0000-0002-0746-1980
Kristin S. Sotzen @ https://orcid.org/0000-0001-7393-2368
Kevin B. Stevenson @ https://orcid.org/0000-0002-7352-7941
Rodrigo Luger @ https://orcid.org/0000-0002-0296-3826
Erin M. May @ https://orcid.org/0000-0002-2739-1465
L. C. Mayorga @ https://orcid.org/0000-0002-4321-4581
Kathleen Mandt @ https://orcid.org/0000-0001-8397-3315
Noam R. Izenberg @ https://orcid.org/0000-0003-1629-6478

References

Agol, E., Dorn, C., Grimm, S. L., et al. 2021, PSJ, 2, 1
Astropy Collaboration, Robitaille, T. P., Tollerud, E., et al. 2013, A&A, 558, A33
Bagnasco, G., Kolm, M., Ferruit, P., et al. 2007, Proc. SPIE, 6902, 69020M
Baranov, Y. I., Lafferty, W. J., & Fraser, G. T. 2004, JMoSp, 228, 432
Barstow, J. K., Aigrain, S., Irwin, P. G. J., Kendrew, S., & Fletcher, L. N. 2016, MNRAS, 458, 2657
Barstow, J. K., Aigrain, S., Irwin, P. G. J., & Sing, D. K. 2017, ApJ, 834, 50
Barstow, J. K., Changeat, Q., Garland, R., et al. 2020, MNRAS, 493, 4884
Batalha, N., Stevenson, K., Hill, M., et al. 2018, natashabatalha
Batalha, N. E., Lewis, N. K., Line, M. R., Valenti, J., & Stevenson, K. 2018, ApJL, 856, L34
Batalha, N. E., Mandell, A., Pontoppidan, K., et al. 2017, PASP, 129, 064501

Lustig-Yaeger et al.
