



Understanding the Effects of Systematics in Exoplanetary Atmospheric Retrievals

Jegug Ih and Eliza M.-R. Kempton

Department of Astronomy University of Maryland, College Park, MD 20742, USA; jegugih@astro.umd.edu

Received 2020 December 16; revised 2021 June 2; accepted 2021 July 21; published 2021 November 11

Abstract

Retrieval of exoplanetary atmospheric properties from their transmission spectra commonly assumes that the errors in the data are Gaussian and independent. However, non-Gaussian noise can occur due to instrumental or stellar systematics and the merging of discrete data sets. We investigate the effect of correlated noise and constrain the potential biases incurred in the retrieved posteriors. We simulate multiple noise instances of synthetic data and perform retrievals to obtain statistics of the goodness of retrieval for varying noise models. We find that correlated noise allows for overfitting the spectrum, thereby yielding a better goodness of fit on average but degrading the overall accuracy of retrievals. In particular, correlated noise can manifest as an apparent non-Rayleigh slope in the optical range, leading to an incorrect estimate of cloud/haze parameters. We also find that higher precision causes correlated results to be further off from the input values in terms of estimated errors. Finally, we show that while correlated noise cannot be reliably distinguished with Hubble Space Telescope observations, inferring its presence and strength may be possible with James Webb Space Telescope observations. As such, we emphasize that caution must be taken in analyzing retrieved posteriors and that estimated parameter uncertainties are best understood as lower limits.

Unified Astronomy Thesaurus concepts: [Exoplanet atmospheres \(487\)](#)

Supporting material: figure sets

1. Introduction

Inverse modeling the transmission spectra of exoplanets allows for the extraction of information about various properties and processes in the atmosphere. This is commonly done by pairing a forward model, which generates a spectrum from atmospheric parameters, and a parameter estimation scheme, which samples the parameter space to compute the probability distribution of the set of parameters. This method of analyzing observed spectra, called atmospheric retrieval, originally developed for remote sensing in solar system bodies, (e.g., Rodgers 2000) has become a standard tool in characterizing exoplanetary atmospheres and has allowed for the measurement of abundances of various species and the identification of atmospheric phenomena such as the presence of clouds and thermal inversions (e.g., Line et al. 2013; Waldmann et al. 2015; Madhusudhan 2018; Benneke et al. 2019; Zhang et al. 2019; Barstow & Heng 2020).

Recently, there has been a growing body of work that addresses the potential to be misled by incomplete physics or simplifying assumptions used in retrievals, often invoked to speed up the computation and make the retrieval feasible (e.g., Pinhas et al. 2018; Caldas et al. 2019; Changeat et al. 2019; Barstow et al. 2020; Lacy & Burrows 2020; Taylor et al. 2020). These studies are especially germane in preparation for interpreting James Webb Space Telescope (JWST) observations, the precision of which will now render the finer details of the model consequential. Such details include three-dimensional atmospheric structure, host star effects, aerosol models, and disequilibrium chemistry. The general method used in the aforementioned works is to retrieve on a synthetic spectrum generated by a more complex and complete forward model and then inspect the retrieval result to quantify what bias may be incurred, with the aim of making judicious choices as to how and which model complexity and compromises should be introduced to the retrieval’s forward model.

Adopting a similar approach, we focus on a separate but related issue in this work. A universal assumption made in atmospheric retrieval is that the reported errors in the observed spectrum are Gaussian and independent. This assumption is encoded into the cost function one tries to minimize during the parameter estimation, which is invariably a chi-squared statistic that does not take the covariance between the residuals into account (Andrae et al. 2010; Line et al. 2013; Zhang et al. 2019). While this assumption is a reasonable starting point for analyzing observations, as data quality reaches unprecedented precision and as retrievals incorporate increasingly sophisticated forward models and more rigorous statistical methods, it is necessary to understand the significance of the assumption of independent errors.

As further motivation for this work, there have been observations that hint at the extent to which errors may be correlated with wavelength. To pull one such example from the literature, we identify the observed spectrum of HD 97658b with the Wide-Field Camera 3 (WFC3) instrument on the Hubble Space Telescope (HST; Guo et al. 2020). The retrieval on this data set strongly favors either high-metallicity or cloudy atmospheres, corresponding with a nearly featureless transmission spectrum (a flat line). However, none of the forward models considered by Guo et al. (2020) provide an excellent fit to the data. For example, we show the best-fit PLATON (Zhang et al. 2019) spectrum in Figure 1, which yields a reduced chi-squared of 2.5 and is ruled out by the data at 4.9σ . (We note that Guo et al. 2020 can find models that provide a high-quality fit to their data by scaling their formal error bars by a factor >1 —a practice that we weigh in on later in this paper and that we do not endorse.) As can be seen in Figure 1, the best-fit spectrum produces residuals that are possibly correlated. One rudimentary method of quantifying correlation is to count the number of zero-crossings, which should follow a symmetric binomial distribution if the noise were independent.

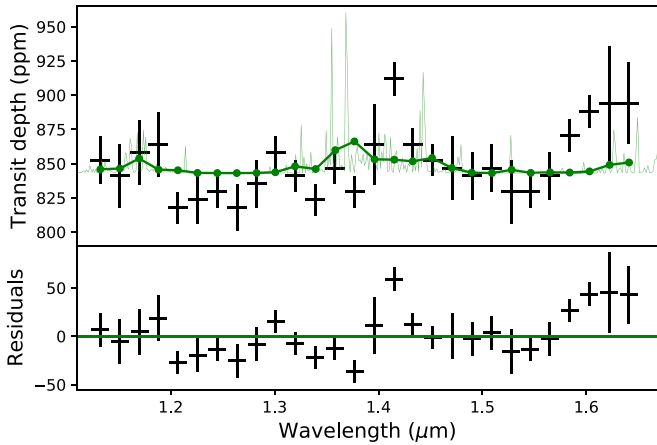


Figure 1. The observed WFC3 data of HD 97658b (Guo et al. 2020) and the best-fit PLATON spectrum in full-resolution and binned (solid line). The residuals from the fit are also shown in the bottom panel. The best-fit χ^2 is 56, and the reduced is $\chi^2_r = 2.5$. An unusual upward trend in the residuals is present in the long wavelength limit, and the residuals appear to be correlated with the wavelength.

(The complete figure set (5 images) is available.)

Additionally, an unusual upward slope is seen in the residual in the red-most edge. A similar behavior was seen in the observed spectrum of KELT-11b (see Colón et al. 2020, their Figure 20). This could be attributed to either complicated physics unaccounted for in the forward model or the presence of correlated noise in the data.

Potential sources of correlated noise are numerous and include faulty data reduction due to incorrect orbital parameters or incomplete subtraction of stellar contributions (Rackham et al. 2018; Bruno et al. 2020). Choices made during the data reduction, such as the choice of a model to remove visit-long trends, can potentially produce a wavelength-dependent correlated effect in the spectrum (see Guo et al. 2020, their Figure 7). More fundamentally, the removal of instrumental systematics such as ramping or horizontal shifts in HST data has intrinsic uncertainties and may potentially manifest themselves in a wavelength-dependent manner (Tsiaras et al. 2016).

For space-based observation facilities, there are some reasons from observer experience to suspect that correlated noise is more likely in the case of a very bright host star, for which the instrumental systematics either behave differently or become more apparent due to lower photon noise. In Stevenson & Fowler (2019, Figure 9), it can be seen that no observations with bright host stars of J -band magnitude ≤ 8.5 meet the ideal precision per orbit, instead maxing out at ~ 35 ppm. (Interestingly, our previously discussed example case of HD 97658b fits into this category, with a J -band magnitude of 6.2.) This effect has been attributed to unaccounted for wavelength-dependent systematics, which have no guaranteed way of being completely modeled out. In the case of ground-based observations, the highly time-dependent telluric contamination may also be a potential source of correlated systematics.

A separate, but related cause of wavelength-dependent correlation in data errors arises when combining data from various instruments to gain a wider wavelength coverage. Each instrument has its own instrumental systematics and data reduction pipeline, leading to distinct noise statistics among the data sets. More fundamentally, these observations are not

simultaneous and are hence subject to differing conditions with respect to stellar and planet variability. An insufficient number of observed transits may admit such variability in the data, even if the desired signal-to-noise ratio (S/N) is formally achieved. Some retrieval analyses have included ad hoc offset parameters, which vertically shift all measured transit depths from one data set by a variable amount, to fit for the discrepancies between data sets, but doing so can induce bias in the estimation of other parameters (Hou Yip et al. 2020).

Another issue arises in how outliers in the observed data are interpreted in a retrieval. After fitting a spectrum to data, the presence of anomalous outliers in the residuals is certainly within the expectation of what can happen, and statistical methods such as bootstrapping, though rarely used in retrieval studies, do offer objective criteria to exclude these outliers. However, the fact that we rely on the best-fit spectrum to determine the outliers raises the question if a statistically equivalent datum could have been accommodated for as a detected feature if it happened to have occurred where we expect one. This problem is especially pertinent in the context of retrievals with nonequilibrium models. It particularly affects resolution-limited observations and species with single, narrow features, e.g., atomic lines such as Na or K for which only one or two data points dictate the retrieved abundance.

Given the number of potential issues raised above, in the present work we aim to address the question: How reliable are our atmospheric retrievals and what are the best practices in the face of these idiosyncratic data systematics? We perform atmospheric retrievals on simulated spectra, while varying the noise properties, and conduct a detailed analysis of the retrieved parameters. Such an analysis provides a statistical context in which one can assess the credibility of a retrieval beyond a raw retrieved posterior. In what follows, in Section 2 we describe our planet parameters and noise models used, in Section 3 we present our findings in how retrievals are affected, in Section 4, we test whether correlation can be estimated during retrieval, in Section 5 we discuss how we might be able to better understand the sources of these systematics and implications for future telescopes, and in Section 6 we summarize and conclude.

2. Methods

2.1. Framework for Statistics of Retrievals

To study how non-Gaussian errors can bias the retrieval results, we perform retrievals on simulated data generated with and without correlation in the noise. Here, by using the same forward model to create the synthetic spectrum and to retrieve it, we remove model dependencies as much as possible and are able to examine the bias due to noise in isolation. To obtain the statistics of retrievals, we use the following procedure (also shown schematically in Figure 2).

1. Choose the input parameters of a planet, such as radius, temperature, and metallicity.
2. Run the forward PLATON (Zhang et al. 2019) model to produce the unpolluted spectrum of the atmosphere.
3. Bin the full-resolution unpolluted spectrum to the chosen wavelength bins.
4. Choose a noise model (independent or correlated) and noise parameters that simulate the noise of a real observation.

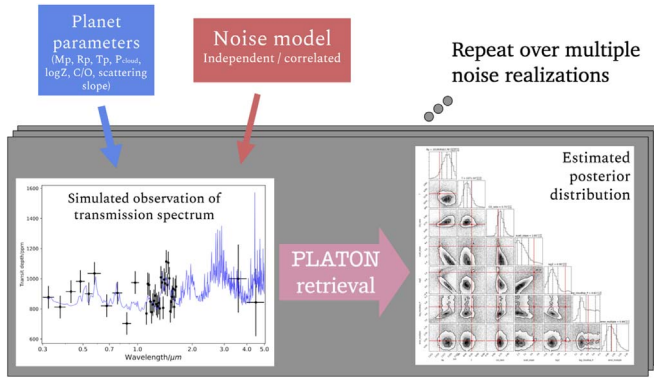


Figure 2. Schematic diagram of our method. We generate multiple (~ 500) observational instances of a given planetary scenario and noise model, and perform atmospheric retrieval on each spectrum using PLATON.

(The complete figure set (5 images) is available.)

5. Sample a noise realization of the binned spectrum using the noise model.
6. Perform retrieval analysis on the simulated data.
7. Repeat steps 5 and 6 a sufficient number of times such that the retrieved parameters can be combined to generate reliable statistics.

We use PLATON (version 3.0; Zhang et al. 2019), an open-source retrieval tool, as the forward and retrieval model for the transmission spectra. PLATON has the advantage of being extremely fast for a retrieval code ($\lesssim 30$ minutes per run), which is suitable for our application as we perform hundreds of retrievals with randomly sampled noise realizations. We perform this process only on transmission spectroscopy, as the geometry allows for the assumption of an isothermal atmosphere, greatly reducing the number of free parameters in our model as well as the computation time per run.

We repeat the above procedure for five cases of observation: a clear hot Jupiter, a clear hot Jupiter with offsets between data sets, a cloudy hot Jupiter, a clear hot Jupiter at higher (James Webb Telescope (JWST)-like) precision, and a warm Neptune. In what follows, we describe the forward model parameters, the noise model, and the retrieval setup. A summary of the input parameters and the assumed priors for each set of retrievals is provided in Table 1.

2.2. Forward Model Parameters

To best imitate retrievals on actual observations, we choose input parameters and spectrum bins similar to HST and Spitzer observations of the canonical hot Jupiter HD 209458b and exo-Neptune GJ436b which have reliable data and have been studied in the context of retrievals (Evans et al. 2015; Deming et al. 2013; Knutson et al. 2007; Deming et al. 2011; Dittmann et al. 2009; Lothringer et al. 2018). We adopt the measured mass, radius, and temperature of each planet, and set the log-metallicity to 0.3 and carbon-to-oxygen ratio to the solar value of 0.53 (Asplund et al. 2009).

We also choose to include cloud and hazes in our model. Such aerosols have been found to be ubiquitous in exoplanetary atmospheres (e.g., Kreidberg et al. 2014; Sing et al. 2016) and produce a spectral signature that can be degenerate with other parameters (Deming & Seager 2017; Marley et al. 2014). PLATON accounts for clouds and hazes using a parametric model. The user can specify a gray cloud-top pressure, the

atmosphere absorption below which is truncated, and an amplitude and slope in the optical end of the spectrum to account for a non-Rayleigh slope caused by Mie scattering. In summary, the aerosol opacity κ_{aer} is given as

$$P > P_{\text{cloud}}: \kappa_{\text{aer}} = \infty$$

$$P < P_{\text{cloud}}: \kappa_{\text{aer}} \propto a\lambda^{-\gamma},$$

where $a = 1$ and $\gamma = 4$ corresponds to Rayleigh scattering from the gaseous atmosphere.

For our cloud-free simulations, we choose a low-altitude cloud-top pressure of 0.5 bar. For our cloudy simulations, we choose 0.1 mbar such that clouds obscure roughly half of the spectrum while preserving some molecular features. Similarly, we adopt a nearly Rayleigh slope of 4.3 and an amplitude of 1, indicating no excess Rayleigh scattering from the aerosol particles. We stress that since we use the same forward model in the retrieval to isolate the effects of noise from model systematics, the specific choices in parameters are not of great importance as long as they can be correctly retrieved and as long as we select a set of forward models that span a representative set of exoplanet atmospheres.

We choose a wavelength range that spans observations from the HST spectrographs most commonly used for exoplanet atmospheric investigations—specifically the Space Telescope Imaging Spectrograph (STIS) and WFC3—as well as photometric observations from the Spitzer Space Telescope. In the STIS wavelength range we follow the bins of Knutson et al. (2007), in the WFC3 wavelength we choose 33 equal sized bins between $1.01 \mu\text{m}$ and $1.64 \mu\text{m}$ (Gennaro 2018), and for Spitzer we include observations in the photometric bands of the IRAC instrument at $3.6 \mu\text{m}$ and $4.5 \mu\text{m}$. The resulting wavelength bins are comparable to a complete panchromatic data set from current space-based observations. The resulting simulated spectra are shown in Figures 3 and 4.

2.3. Noise Model

To simulate the observed data, we sample multiple noise instances centered around the unpolluted spectrum, treating the simulated unpolluted spectra as a multivariate Gaussian distribution with the unpolluted depths as the mean and the reported error at each bin as the width. In addition, we adopt a nondiagonal covariance matrix with an exponential kernel to simulate correlated noise, such that the matrix element that correlates the wavelength bin at λ_i and λ_j is given by

$$K_{ij} = \epsilon_{ij} \sigma_i \sigma_j \exp\left(-\frac{|\lambda_i - \lambda_j|}{l}\right), \quad (1)$$

where σ_i is the reported error at wavelength λ_i , and ϵ_{ij} is 1 for a pair of points from the same data set and 0 otherwise. We choose the scaling factor l to be the distance to the neighboring bin. We select this covariance matrix in particular because it allows for the best-fit spectrum to the WFC3 observations of HD 97658b in Guo et al. (2020) to yield a reduced chi-squared of ~ 1 , as opposed to 2.5 when the errors are construed to be independent and Gaussian. We choose a random noise error of 75 ppm for all instruments, which represents a moderate quality data for STIS and WFC3 but is better than average for typical Spitzer observations (Sing et al. 2016; Garhart et al. 2020). We also assume a uniform transmission across the wavelength range of each Spitzer filter in both the forward and retrieval models. In practice, we find that the two broadband Spitzer points provide little constraint on the retrieved parameters,

Table 1
Summary of Equilibrium Chemistry Retrievals and the Input Parameters and Priors Used

Parameter	Name	Clear HJ		Clear HJ with Offset	
		Truth Value	Prior	Truth Value	Prior
M_p	Planetary mass (M_J)	0.73	15%	0.73	15%
R_p	Planetary radius (R_J)	1.42	15%	1.42	15%
T	Temperature (K)	1130	[300, 1500]	1130	[300, 1500]
C/O	Carbon-to-oxygen ratio	0.53	[0.2, 2.0]	0.53	[0.2, 2.0]
$\log a$	Log-scattering factor	0	[−0.3, 0.3]	0	[−0.3, 0.3]
γ	Scattering slope	4.3	[2.0, 5.5]	4.3	[2.0, 5.5]
$\log Z$	Log-metallicity	0.3	[−1, 3]	0.3	[−1, 3]
$\log P_{\text{cloud}}$	Cloud-top pressure (log(Pa))	4.7	[2.5, 6.5]	4.7	[2.5, 6.5]
Offset 1	STIS offset (ppm)	...	...	−100	[−300, 300]
Offset 2	Spitzer offset (ppm)	...	...	150	[−300, 300]
Error multiple	...	Unity	[0.5, 2.0]	Unity	[0.5, 2.0]
<i>Other parameters</i>					
R_s	Stellar radius ($R_\odot$)	1.19		1.19	
T_s	Stellar temperature (K)	6090		6090	
Data error	(ppm)	75		75	
# of runs		440		660	

Parameter	Truth Value	Cloudy HJ		Clear HJ, High-precision		Warm Neptune	
		Truth Value	Prior	Truth Value	Prior	Truth Value	Prior
M_p	0.73		15%	0.73	15%	0.0736	15%
R_p	1.42		15%	1.42	15%	0.395	15%
T	1130		[300, 1500]	1130	[300, 1500]	700	[300, 1500]
C/O	0.53		[0.2, 2.0]	0.53	[0.2, 2.0]	0.53	[0.2, 2.0]
$\log a$	0		[−3.0, 3.0]	0	[−0.3, 0.3]	0	[−2, 2]
γ	4.3		[2.0, 5.5]	4.3	[2.0, 5.5]	4.3	[,]
$\log Z$	0.3		[−1, 3]	0.3	[−1, 3]	0.3	[−1, 3]
$\log P_{\text{cloud}}$	3.0		[0.0, 6.5]	4.7	[2.5, 6.5]	5.0	[2.5, 6.5]
Offset1	...		...	...	...	...	...
Offset2	...		...	...	...	...	...
Error multiple	Unity		[0.5, 2.0]	Unity	[0.5, 2.0]	Unity	[0.5, 2.0]
<i>Other parameters</i>							
R_s	1.19			1.19		0.683	
T_s	6090			6090		4780	
Data error	75			10		75	
# of runs	660			400		400	

Note. The stellar parameters shown are not retrieved for. For the mass and radius, the priors show the width of the Gaussian prior, relative to the input value. For other parameters, the priors are uniform in the interval. For log values, the priors are log-uniform in the interval.

and using the same error for all instruments does not give undue importance to the Spitzer points. For the high-precision hot Jupiter case, we use 10 ppm errors to match the best current data quality (Colón et al. 2020).

We show a few randomly selected noise realizations in Figure 5 for the Gaussian and correlated noise models. It is discernible from the bottom row of panels that the correlated noise has slightly redder residuals compared to the Gaussian noise; that is, there are less zero-crossings as the neighboring points are correlated. We also stress that overall this is a rather subtle effect; without knowing the unpolluted depths a priori to produce the residuals, from the spectrum alone it is hardly obvious that there is a distinction between the two.

Additionally, to examine the effects of including the offset between data sets as a retrieved parameter, we create spectra with and without a fixed offset between the data sets. Namely, we add a vertical offset of −100 ppm to the STIS points and +150 ppm to Spitzer points, holding the WFC3 points

constant. The specific amount of offset is a somewhat arbitrary choice. The relevant heuristic is that the offset should be exactly retrieved in the absence of degeneracy.

2.4. Retrieval Setup

Using PLATON, setting up the retrieval involves choosing the priors and use of the statistical sampling method. We choose the priors to be comparable to a real retrieval analysis. PLATON accepts either a Gaussian prior or a uniform prior in a user-specified interval. We set Gaussian priors for the planet radius and mass, as these are often constrained via other methods such as radial velocity and transit measurements prior to observing the transmission spectrum. The Gaussian prior is centered around the input value and with a standard error of 15%. This precision is comparable to or slightly overestimates the typical uncertainty in the measurement of mass via the radial velocity method and provides sufficient tolerance for the retrieved value to deviate from the input value, if necessary. For the instrumental offsets, to ensure

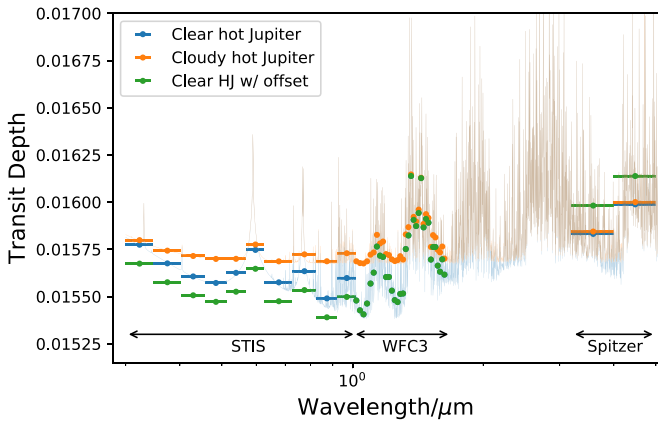


Figure 3. Simulated unpolluted spectra of the hot Jupiter cases. Both full-resolution and binned depths are plotted. The offset case is identical to the clear case in the WFC3 band. The main effect of the clouds is the truncation of the bottommost depths compared to the clear case.

that the prior is broad enough, we choose a uniform prior offset to be two- and three-fold of the offsets. For all other parameters we opt for uniform priors that are as wide and uninformative as possible and adopt the full parameter range supported.

For now, we only choose to do retrievals with equilibrium chemistry, where the composition of the atmosphere at a given temperature and pressure is dictated by the global elemental abundances set by metallicity and the carbon-to-oxygen ratio. While disequilibrium chemistry is indeed expected for planets below $T_{\text{eq}} \leq 1200\text{K}$, most of its effects take place below the altitude that is typically probed by transmission spectroscopy and have no easily discernible effect on the spectrum at the data precision of current instruments. The actual discrepancy in the relevant pressure range (~ 1 mbar) is smaller than the uncertainties we can currently obtain (e.g., Line et al. 2011; Miller-Ricci Kempton et al. 2012; Fortney et al. 2020).

PLATON supports either Markov chain Monte Carlo (MCMC; Foreman-Mackey et al. 2013) or nested sampling methods for the posterior estimation (Feroz & Hobson 2008). We note that while both are statistically robust and widely used, we observe a minor discrepancy in the resulting posteriors between the two methods, in which the posteriors estimated by MCMC tend to be slightly broader than those estimated by nested sampling. This does not pose a major issue for this work inasmuch as we are concerned with biases due to data idiosyncrasies, and we consistently use one algorithm across our analyses. Nevertheless, we draw attention to this point as it requires extra scrutiny when comparing retrieval results. We choose nested sampling as it is known to perform better in estimating multimodal or oddly shaped distributions.

3. Results

In this section we first present the overall effects of correlated systematics on our retrievals, using the clear hot Jupiter case as a baseline. We then examine which parameters in particular are affected. We finally show how the results for the baseline case also extend to the other retrieval cases, and point out additional effects that arise.

3.1. Overfitting Due to Correlated Noise

Here we present the effects of correlation in data on the retrieval overall. To do this, we must reduce a retrieval result

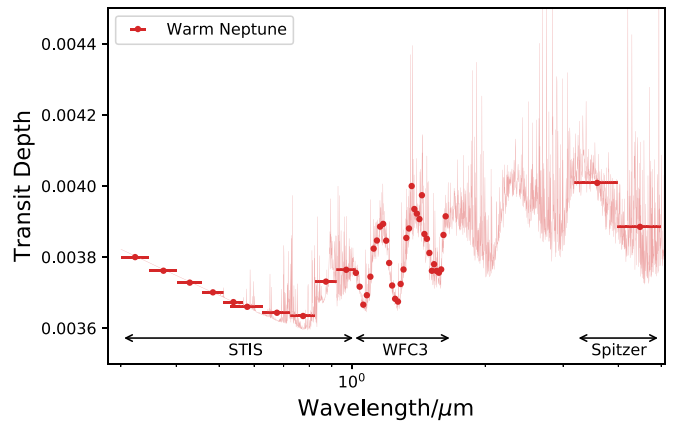


Figure 4. Same as Figure 3, but for the warm Neptune case.

into a simpler metric. In retrievals on actual data this is commonly done by using the reduced chi-squared between the observed data and the best-fit or median spectrum. Here, as we begin from a known simulated ground truth, we can also compare the retrieved posterior directly to the input values to measure the accuracy of the retrieval by using a probability integral transform (PIT), which is the cumulative distribution function evaluated at the input value. To do this, in each retrieved posterior distribution, we draw an isolikelihood contour of the input parameters and sum the relative weights contained within, producing a confidence interval between 0 (the input was the most likely sample) and 1 (the least likely). The distribution of the PIT values should follow a uniform $\mathcal{U}(0, 1)$ were the retrievals accurate.

Our main finding is that, on average, correlation in the data allows for overfitting the spectrum, thereby weakening the overall accuracy of the retrieval. In other words, the best-fit spectrum is more likely to achieve a reduced chi-squared lower than unity in the presence of correlated noise, while simultaneously the retrieved posterior distribution rules out the input with higher significance.

In Figure 6, we present the comparison between Gaussian and correlated noise for the clear hot Jupiter case. For both retrievals with Gaussian (black) and correlated noise (blue), we show the distributions of: χ^2 between the unpolluted spectrum and the data instances, the reduced chi-squared between the best-fit spectrum and the data instances, and the PIT values showing the accuracy of retrieval, as described above. A few observations can be made:

1. The fit χ_r^2 , or the goodness of fit of the retrieval, on synthetic data is on average skewed to better than unity in the presence of correlated noise. In other words, it is more likely that the retrieval will overfit the data with forward models.
2. The accuracy of the retrieval, shown as the PIT value, on the other hand, is *worse* in the presence of correlated noise. We also show the cumulative distribution in Figure 7 to demonstrate the worsening of the accuracy. The Kolmogorov–Smirnov (K-S) statistic between the two cases is 0.19.
3. For the normal noise case, even in the absence of the correlated noise, the retrieval accuracy is close to but slightly worse than the expected uniform distribution. This minor discrepancy is likely due to the fact that the retrieved posterior distribution is already non-Gaussian

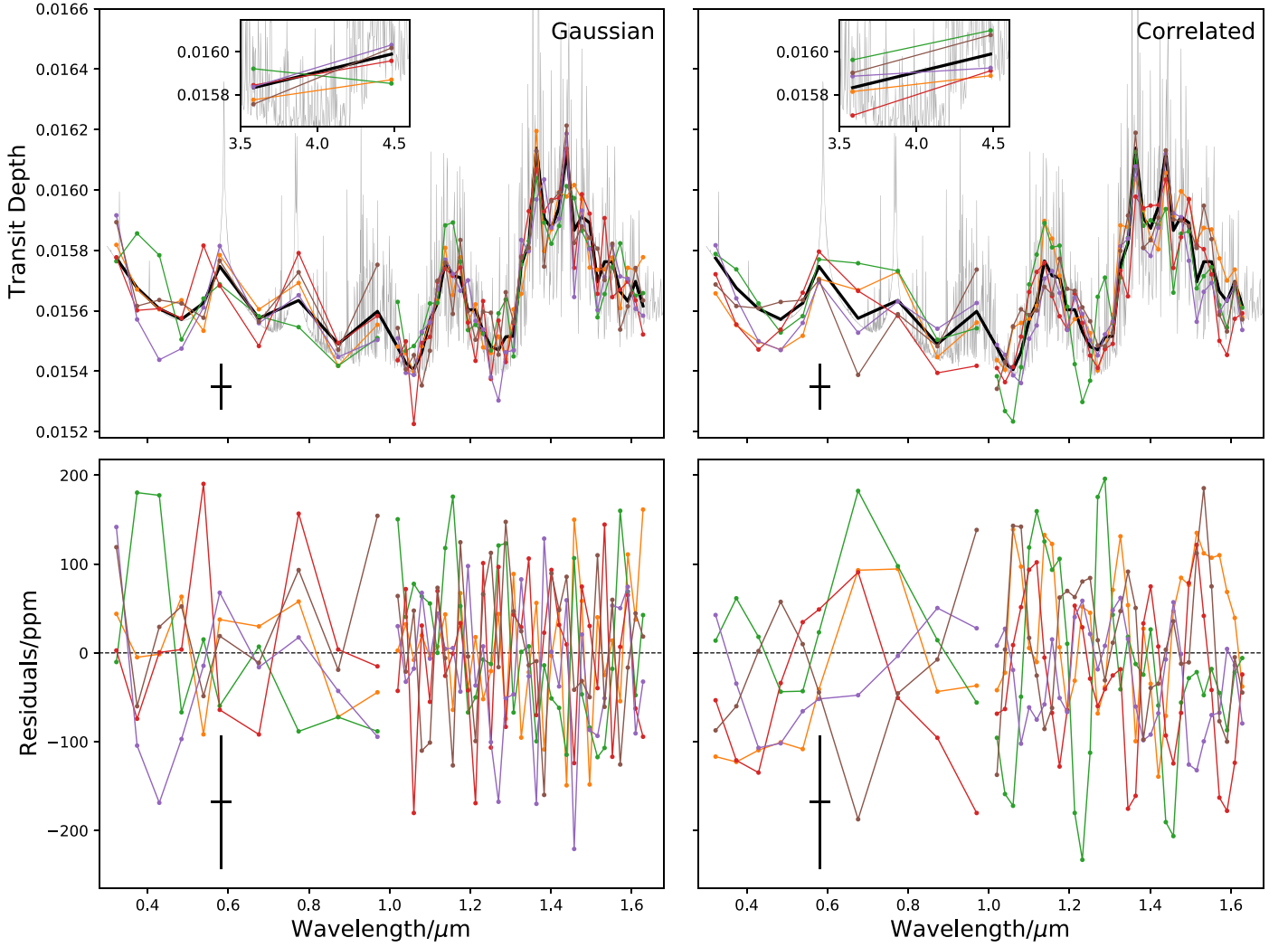


Figure 5. Comparison of Gaussian (left column) and correlated (right column) noise models for the clear hot Jupiter case. In the top row of the panels, five randomly selected spectrum realizations are plotted in color and the unpolluted spectrum is plotted in black. The disconnection in the lines indicates discrete instruments. The Spitzer data are shown separately in the inset plot. The assumed 75 ppm error bar is shown for scale. In the bottom row of panels, the residuals relative to the unpolluted spectrum are shown. The effect of wavelength correlation is discernible in the slightly redder noise (i.e., the residuals appear subtly sparser due to less zero-crossings as neighboring residuals are more likely to have the same sign) in the bottom right panel, but would not be discernible in the top right panel without prior knowledge of the ground-truth spectrum.

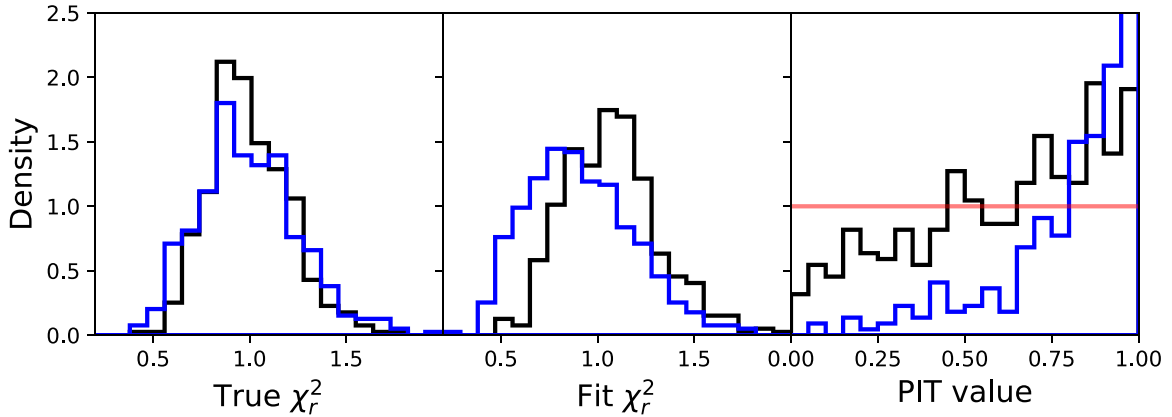


Figure 6. Two-dimensional distribution comparing the retrievals with Gaussian noise (black) and correlated noise (blue). The three parameters shown are: the chi-squared between the unpolluted spectrum and the observation instance (true χ^2), the reduced chi-squared between the best-fit spectrum and the observation instance (fit χ^2), and the PIT values for the retrieved posterior. The correlated noise allows for overfitting the spectrum, while simultaneously degrading the accuracy of the retrieval. We remind the reader that this corner plot is not showing a single retrieved posterior result, but a composite of multiple posteriors.

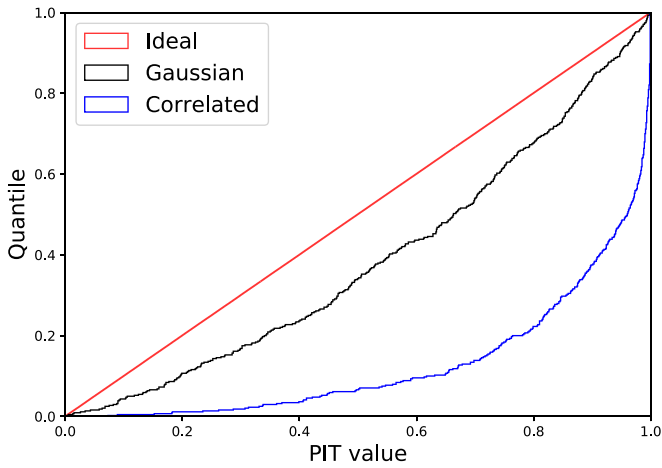


Figure 7. Same distribution as in the right panel of Figure 6, but shown as a cumulative distribution. The horizontal axis corresponds to the PIT values for each retrieval; the vertical axis corresponds to the cumulative fraction of all retrievals. The K-S metric measures the maximum vertical discrepancy between two cumulative distributions, such as above, and values of this metric are reported in Table 2. The expected uniform distribution (red), and our independent noise (black) and correlated noise (blue) simulation cases are shown.

for a few parameters (discussed in Section 3.2), as well as due to degeneracy between certain parameters.

4. While the goodness of fit is on average better in the presence of correlated noise, it is not the case that the distributions of fit χ_r^2 are so discrepant that one can deduce the presence of correlated noise from the goodness of fit alone. That is, a given overfit spectrum can plausibly be construed as either a consequence of correlated noise or as an unlucky instance of Gaussian noise that happens to lie at the tail of the χ_r^2 distribution. As such, we stress that the effect of correlated noise is manifested statistically, and no individual value of χ_r^2 , good or bad, is uniquely a diagnostic of correlated noise in a single retrieval instance.

3.2. Which Parameters Are Affected?

Given that the correlated noise degrades the overall accuracy of retrievals, it is necessary to then look at which parameters are affected. We do this by marginalizing the posterior distribution over each parameter, as is typically done in retrieval analyses. The effect on each marginalized distribution can be twofold—the mean can be biased, the estimated error can be affected, or both. Either a shift in the mean away from the input parameter or an underestimation of the error can worsen the accuracy of the retrieval. As such, we examine the retrieved mean and the retrieved error separately for each parameter.

The distribution of the retrieved means is shown in Figure 8. For the case in which noise is independent (black), the retrieved means form clean normal distributions around the input values (red) for most parameters, as expected from the central limit theorem. The two exceptions are the C/O ratio and the cloud-top pressure. This is most likely due to the fact that the retrieved distributions for these parameters are not Gaussian in the first place. For cloud-top pressure, the retrieved distribution is at best a flat distribution with a lower bound, ruling out a cloudy atmosphere as per the clear atmosphere in the input used. For the C/O ratio, we suspect that the distribution is skewed due to the increasing influence the

parameter has over its range. That is, as one sweeps through the C/O ratio, the spectrum changes more rapidly over the range above the solar value of 0.53, and thus the means are naturally skewed to values lower than the solar value where there is a greater density of near-consistent solutions.

The presence of correlated noise has a few interesting effects on the retrieved means. First off, the error multiple parameter is biased to less than unity. This intuitively follows from the global result that correlated noise allows for overfitting of the spectrum, tricking the error multiple parameter to believe that the error bars are overestimated. This means that the error multiple parameter is more likely to behave pathologically in a situation where one may expect it to be useful, such as if the reported error bars truly were underestimated due to unknown and unaccounted systematics. The retrieval instead selects a less-than-unity value of the error multiple, incorrectly implying that the data precision is better than initially reported. This is possible if the domain of input parameters and the forward model can still reproduce the spectrum polluted with systematics.

In the correlated noise case, the retrieved means generally show a wider distribution to varying degrees for each parameter. Specifically, the radius, mass, and temperature are the most affected, while the effect is the least pronounced for metallicity and the cloud-top pressure. This result may be explained by considering the wavelength scale the former three parameters have on the spectrum. Mass and temperature affect the scale height of the atmosphere, which affects the overall vertical extent of the transmission spectrum. The radius affects the baseline transit depth as well as the scale height. These are global parameters in the sense that the transit depths in all of the bins are affected together. As such, a wavelength-dependent correlation can bias these parameters. On the other hand, metallicity, while it also affects the scale height (via the mean molecular weight), directly controls the individual transit depths. This has a more local effect in that it changes the actual shape of the spectrum.

The distribution of the retrieved errors is shown in Figure 9. The effect of the correlated noise is clearly visible for all of the parameters in that the retrieved error bars show a tendency to be underestimated. For instance, the retrieved error on log-metallicity is on average underestimated by ~ 0.2 dex. While this disparity is smaller still than typical constraints, it is worth bearing in mind as this is a statistical effect; the spread over the retrieved error is by itself broad enough that the actual effect of a given instance can be much larger than this value. Additionally, when JWST allows for the precise measurement of metallicity, this level of uncertainty may not be negligible when one considers analyzing archival data simultaneously. The same consideration applies to other parameters. As such, in this context we suggest that the retrieved constraints for parameters, in the face of the potential for correlated noise, are best understood as lower limits.

3.3. Extensions to Other Planet Parameters

In this section, we present our results for planet scenarios other than the baseline clear hot Jupiter case to understand the sensitivities of our results to various system and data set parameters. We show histograms of the retrieved mean and retrieved standard error for each parameter in figure sets for the remaining planetary scenarios (the hot Jupiter with offsets, cloudy hot Jupiter, high-precision hot Jupiter, and warm Neptune). We generally find that the main results stated so far hold true for all cases: correlated noise causes both overfitting in χ_r^2 and worsening of the accuracy of retrieval (i.e., larger PIT values).

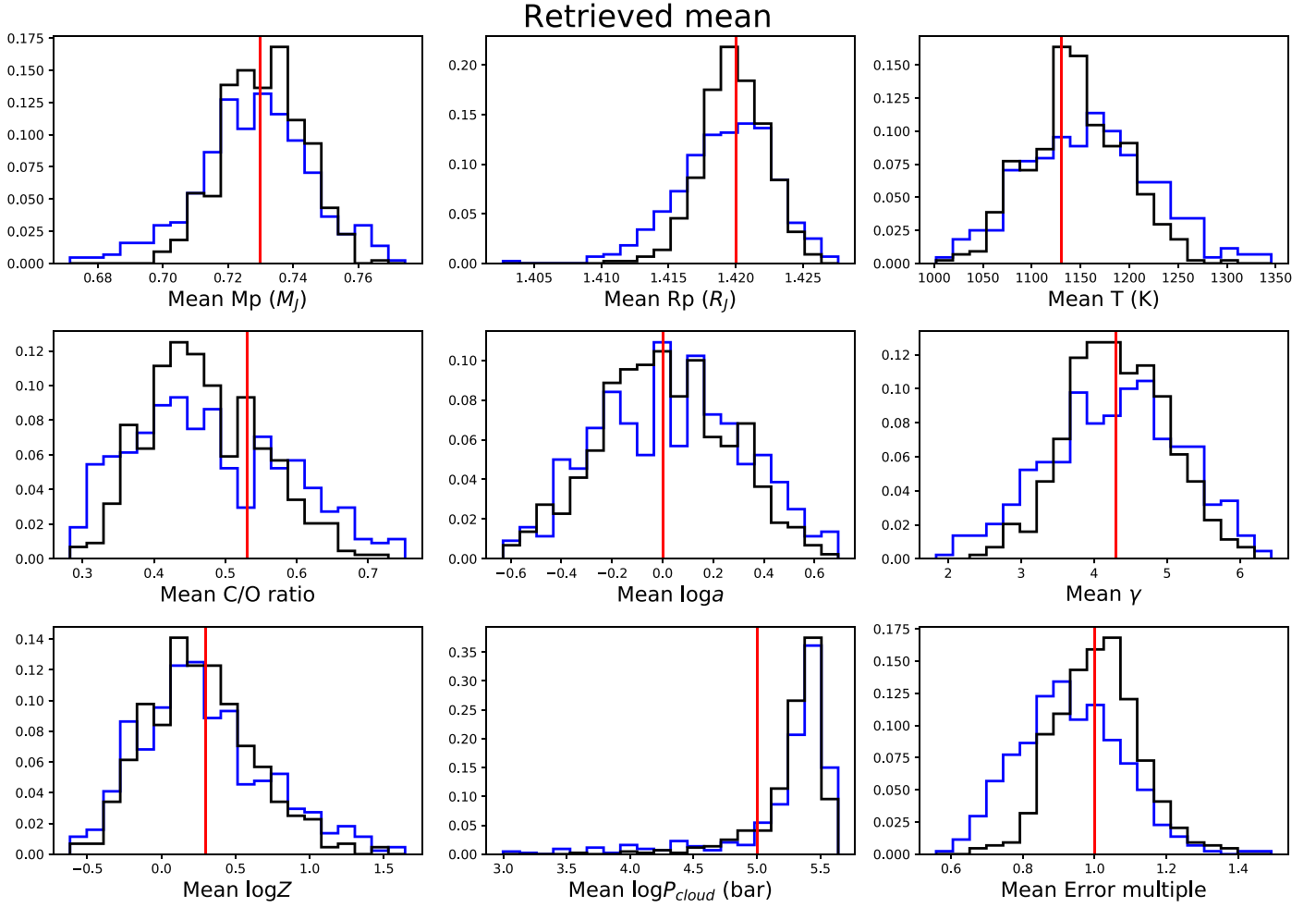


Figure 8. The distribution of retrieved means per each parameter, for independent noise (black) and for correlated noise (blue), for the baseline hot Jupiter case. The input values used to generate the spectrum are shown in red.

This point is summarized in Figure 10, in which we show the medians of the χ_r^2 and PIT value distributions for each planet realization, i.e., distilling down the results of Figure 6 and the like to values quantifying the peak and the spread.

To further quantify this point, we perform a K-S test for the goodness of fit and the accuracy-of-retrieval metrics to measure the discrepancy between the results of Gaussian and correlated noise. In Table 2, we show the K-S statistics, D , for the fit χ_r^2 and the PIT value representing the accuracy of retrieval. We find that the clear hot Jupiter happens to be the best case for the smallest discrepancy of accuracy of retrieval between Gaussian and correlated noise, and that other cases generally result in further discrepancy between results with Gaussian and correlated noise.

3.3.1. Bias in the Non-Rayleigh Scattering Slope

In the retrievals of the hot Jupiter with instrumental offsets and the warm Neptune, we find that the retrieved haze properties also show the potential to be biased. Correlated noise can bias the scattering slope, γ , away from the Rayleigh value of 4, misleading the retrieval to infer the presence of aerosols. This bias makes intuitive sense as, if a handful of points in the optical wavelengths align due to correlated noise, those points can mimic the behavior of a non-Rayleigh slope (May et al. 2020). As such, we caution that a spurious detection of haze

can be possible in interpreting data in which the presence of correlated noise is either expected or suspected.

We suspect that this bias happens more readily for the warm Neptune case compared to the hot Jupiter retrievals because the overall signal is smaller while the data error used to scramble the spectrum was held constant at 75 ppm, resulting in a larger relative error. The warm Neptune spectra consequently have greater potential for large (apparent) optical slopes to manifest.

3.3.2. Retrieving Offsets

We ran a set of retrievals that includes nonzero offset parameters between data sets from different instruments. We find that while the presence of correlated noise does cause underestimation of the uncertainty in the offset in an identical manner to other parameters, it does not worsen the retrieval of the means. The offsets are accurately retrieved in both Gaussian and correlated noise retrievals and do not pose any obvious degeneracies.

This is somewhat surprising as, in our formulation of correlated noise, offsets can be regarded as correlated noise with high correlation and long wavelength order. For instance, in Figure 5, the data instances with correlated noise in the Spitzer band mimics the presence of an offset.

It should be obvious that the influence of offset data points will strongly depend on the specific wavelength those points

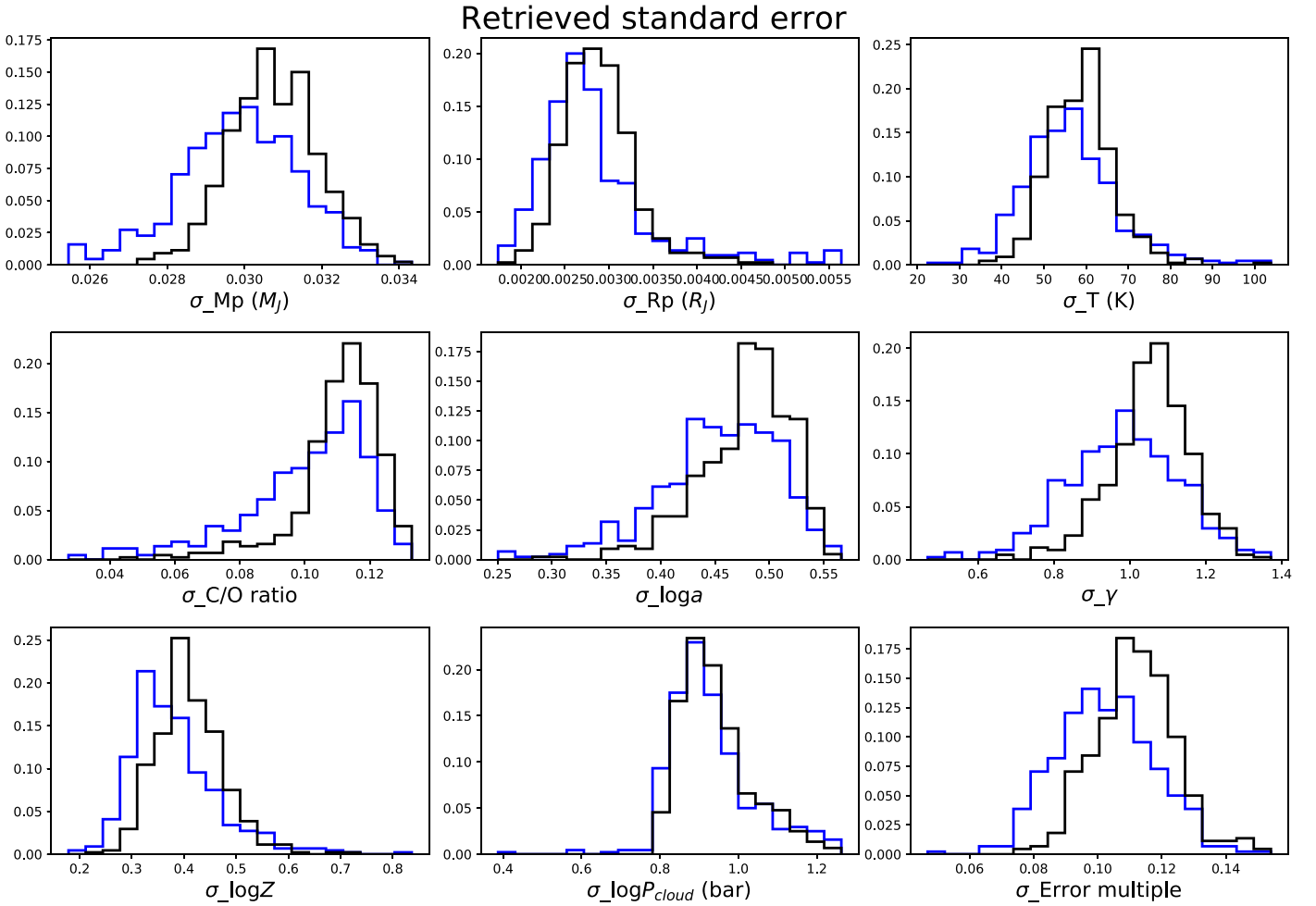


Figure 9. The distribution of retrieved standard errors per each parameter, for independent noise (black) and for correlated noise (blue), for the baseline hot Jupiter case.

occupy, as well as the offset magnitude and sign. As such, we present here only one possible manifestation of how real data could behave. For instance, we have only considered offsets between data sets disjointed in wavelength, but, say, merging data from ground-based and space-based observations can produce offset data with overlapping wavelength coverage. Hou Yip et al. (2020) found that if there is overlapping data with nonzero offsets and if free retrievals are used, such offsets can be degenerate with the estimated abundances if equilibrium chemistry is not assumed.

3.3.3. Effects of Clouds

The main effect of adding gray clouds to the model, from the point of view of the retrieval, is washing out information contained in the spectrum that originates from the high-pressure portion of the atmosphere. In Figure 3, roughly half of the HST points are covered by clouds, no longer constraining, say, a baseline radius or metallicity. We find that the broad effect of underestimating uncertainty and biasing means due to correlated noise still holds for cloudy hot Jupiter retrievals.

For the retrieved cloud-top pressure parameter, the main effect is a bias in the retrieved mean. Specifically, the presence of correlated noise disrupts the distribution of retrieved means of cloud-top pressure by extending the tail in the high-pressure direction. In other words, the spectrum is more likely to be

understood as having a clear atmosphere. Upon examining the spectra for the retrievals that populate this tail, we find that the correlated noise happens to manifest as a number of data points dipping under the opaque cloud top where the atmosphere is normally optically thick, thereby mimicking the behavior of a clear atmosphere.

3.3.4. Effects of Higher Precision

We find that in the hot Jupiter retrieval with high-precision (10 ppm) data, the broad conclusions again still hold. Correlated noise leads to an underestimation of retrieved uncertainty for all parameters. Compared to other cases however, correlated noise does not shift the retrieved means as much, which is to be expected since every noise instance only has a minor deviation from the unpolluted spectrum (specifically a factor of 7.5 times smaller than in our baseline case), even with correlation.

Comparing the high-precision case to the baseline case with 75 ppm errors, we find that, naturally, both the estimated means are retrieved closer to the input values and the retrieved parameter uncertainties are concurrently smaller. Interestingly, the uncertainties shrink more than the means approach the input values; consequently, in the high-precision case, the retrieval more readily rules out the input. This is shown in Figure 11, in which the retrieved means are normalized by their retrieved

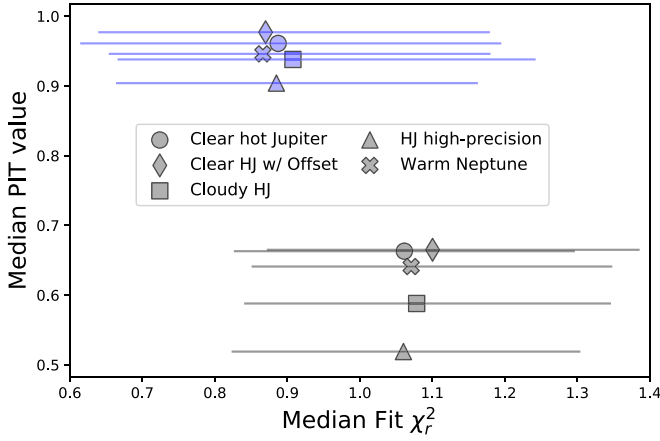


Figure 10. Summary of the goodness of fit and retrieval accuracy for the five planetary cases; black is for Gaussian noise, and blue is for correlated noise. The filled symbols mark the peak in both distributions, and the error bars denote the 1σ spread. Values of χ_r^2 closer to unity imply a better fit. Similarly, lower PIT to a more accurate retrieval result. This plot shows the information in bottom middle panel of Figure 6 for all five cases. For the clear hot Jupiter with offset, the retrieval accuracy is measured as the error in an 11-parameter Gaussian distribution; the rest are quantified in 9 parameters. The correlated retrievals clearly have distinct distributions of goodness of fit and accuracy of retrieval, but have enough overlap such that one cannot discern whether a single retrieval instance has correlated noise.

uncertainty to show the number of standard errors within which the input value for each parameter is retrieved. The high-precision case (dashed line) actually has more retrievals *further* from the input value when normalized. We find an identical trend for the retrievals with correlated noise.

Additionally, we find the PIT value distribution for the retrievals with Gaussian noise much more successfully follows the ideally expected uniform distribution, with a K-S statistic of $D = 0.04$ (see Table 2 and Figure 12). Given that the accuracy of the retrievals for all other cases (with 75 ppm error bars instead) are at least somewhat discrepant from the expected distribution even for the Gaussian noise retrievals, this result suggests that, even for hot Jupiters, 75 ppm error bars are too large to assume a priori that the retrieved posterior will follow a multivariate Gaussian. This has implications for parameter estimation methods that need this assumption of a Gaussian posterior, such as optimal estimation or some of recent machine learning-based retrievals (Line et al. 2013; Cobb et al. 2019). These methods require that the data uncertainty is small enough such that the forward model behaves linearly over the parameter uncertainties. Retrievals using these methods must be trusted only when the data has an exceptional signal-to-noise ratio.

4. Can We Tell If Systematics Are Present?

In the motivating example spectrum of HD 97658b in Section 1, the presence of correlated noise was suspected based on the fact that no forward model can produce a satisfactory fit under the assumption of randomly scattered residuals. If we are to presume that the retrieval is indeed correct, and the residuals evidence correlated noise contaminating a genuine featureless spectrum, then we should also consider how prevalent unnoticed correlated noise can be in the observed data of other planets. The natural question then is to ask whether there is a more robust and comprehensive way of distinguishing correlated noise within the framework of a retrieval. Especially,

Table 2
K-S Statistic of χ_r^2 and PIT Values

Case	χ_r^2 G-C	PIT value		
		U-G	U-C	G-C
Clear hot Jupiter	0.28	0.19	0.60	0.47
Clear hot Jupiter w/offset	0.36	0.19	0.67	0.58
Cloudy hot Jupiter	0.27	0.12	0.54	0.48
Clear hot Jupiter, high precision	0.28	0.04	0.48	0.47
Warm Neptune	0.35	0.16	0.58	0.52

Note. The K-S statistic, D , measures the maximum vertical discrepancy between the cumulative distributions (see Figure 7), of the goodness of fit, and the retrieval accuracy for each planet scenario. The first column shows the two-sample D between the distributions of fit χ_r^2 for the Gaussian and correlated noise. The next two columns contain the D between the expected uniform distribution and the PIT value distributions from retrievals with Gaussian and correlated noise, respectively. The final column shows the D between the two distributions. In all cases, the discrepancy is due to overfitting and worsening of retrievals.

given that correlated noise can give rise to overfitting, it is of special interest whether correlated noise can be distinguished from merely overestimated error bars. A natural way to achieve this is to modify the likelihood function such that it can reward or penalize when the residuals are correlated.

To test this, we implement a parameterized covariance matrix and let the retrieval estimate the hyperparameter that measures the correlation strength. We use a nearest-neighbor correlated noise model as in Sivia & Skilling (1996), where the correlation strength is parameterized by ϵ , such that the covariance matrix element between points at i and j is given by

$$K_{ij} = \sigma_i \sigma_j \epsilon^{|i-j|}. \quad (2)$$

This differs slightly from the correlated noise model used in Section 2 in that this model does not depend on the wavelength difference between two points but depends instead on the difference in indices. The two implementations would be identical if the wavelength grid was regularly spaced, with the exponential base giving a correlation strength of $\epsilon = e^{-1} \sim 0.37$. While an extension to accommodate wavelength-dependent correlation is certainly possible, as a first test this simplification provides a reasonable starting point for exploring whether correlated noise can be retrieved.

This simplification allows one to write the likelihood function as

$$\ln \mathcal{L} = -\frac{1}{2} \left[(N-1) \ln(1 - \epsilon^2) + \sum_{i=1}^N \ln 2\pi\sigma_i^2 + \frac{Q}{1 - \epsilon^2} \right], \quad (3)$$

in which Q is the modified chi-squared-like term related to the error-scaled residuals R_i by

$$Q = (1 + \epsilon^2) \sum_{i=1}^N R_i^2 - \epsilon^2 (R_1^2 + R_N^2) - 2\epsilon \sum_{i=1}^{N-1} R_i R_{i+1}. \quad (4)$$

We perform a small grid of retrievals to study when the correlated noise can be distinguished from overestimated error bars and correctly retrieved. The input parameters for the planet remain the same as Case 1 in Section 2, while we vary the noise properties. Our grid consists of three values of error multiple: $\eta = 0.8, 1$, and 1.25 ; three values of correlation strength: $\epsilon = 0$,

Normalized retrieved parameter

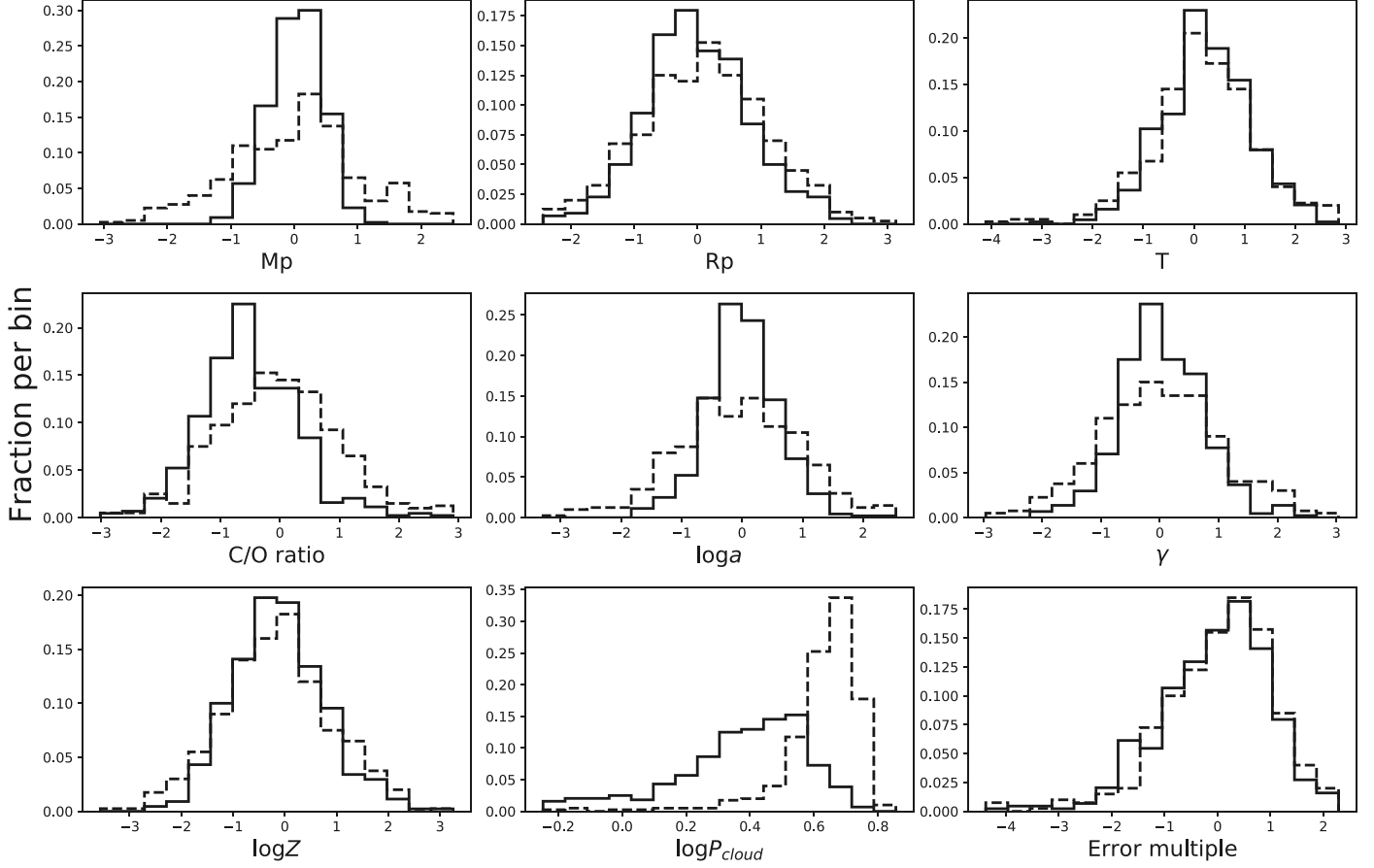


Figure 11. Distributions of each parameter, normalized by the retrieved uncertainty, for the hot Jupiter retrievals. A value of zero indicates that the retrieved mean coincides with the input value. The dashed lines show the high-precision (10 ppm) case, and the solid line shows the baseline (75 ppm) case. Only the Gaussian noise retrievals are shown here, but the correlated noise case displays similar behavior.

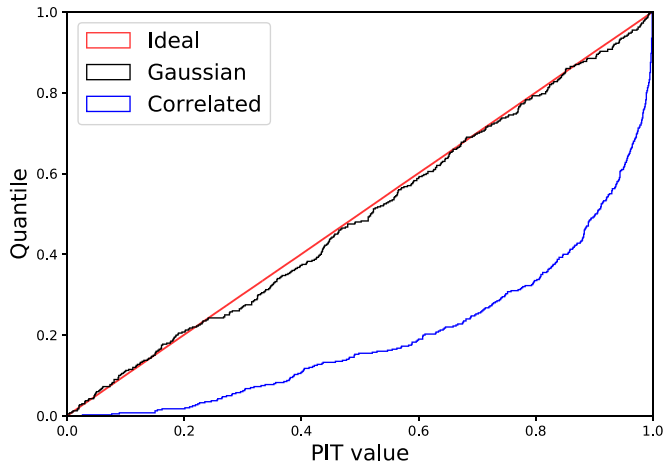


Figure 12. Same as Figure 7, but for the high-precision case. The smaller data error results in a better agreement between the ideal uniform distribution and the retrieval accuracy in the Gaussian noise retrievals.

0.25, and 0.5; and two wavelength grids modeling the HST WFC3 and JWST Near InfraRed Spectrograph (NIRSPEC) prism grids. Our NIRSPEC grid consists of 133 linearly spaced points between 0.6 and 5 microns. From a few initial tests we find that varying the absolute size of the error bar does not change the results, and hence we keep them fixed at 75 ppm.

For each combination of noise parameters, we run three retrievals: one in which both error multiple and correlation strength were included as retrieved parameters, and two in which either was removed. We set a uniform prior between 0 and 1 for the correlation strength, as we do not expect an anticorrelation between neighboring points.

We show the marginalized posterior distribution for correlation strength in Figures 13 and 14 for the WFC3 and NIRSPEC grids, respectively. From the posterior distributions it is clear that the correlation strength can be adequately retrieved, and the posterior width is dependent on the number of points in the data, as expected. It can also be seen that underestimated error bars, when unaccounted for, can be mistaken for the presence of correlated noise, and vice versa, as demonstrated in the previous section. The main difference between the two instruments from the point of view of the retrieval is simply the number of points.

Additionally, we calculate the Bayes factor among the retrievals to determine whether the inclusion of each parameter is warranted. The results for WFC3- and JWST-like data are shown in Figures 15 and 16. The case in which both parameters were included is used as the baseline, and the log-ratio (in base 10) of the Bayesian evidence is shown for each case in the grid. A positive value indicates that the model better fits the data while spanning a smaller prior volume, indicating that the parameter should be removed.

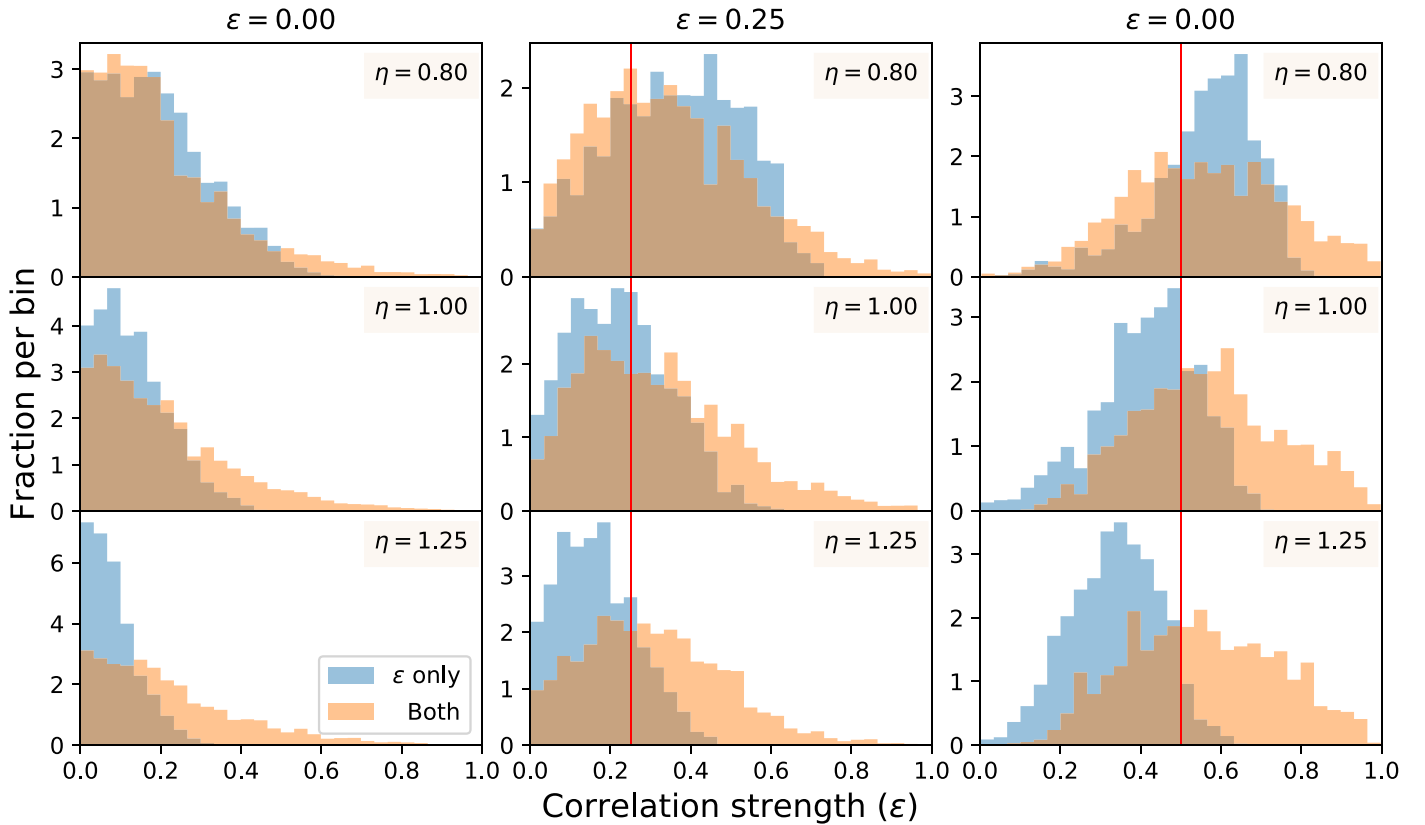


Figure 13. Histograms showing marginalized posterior distribution for the correlation strength parameter for a hot Jupiter on a WFC3-like wavelength grid. The relevant input values for error multiple and correlation strength for each run is shown in the legend. Two posteriors are shown for each: one where only the correlation strength was retrieved (blue) and one where both the error multiple and correlation strength were retrieved (orange). The corresponding marginalized posteriors for the error multiple are not shown. We note that our original model in Section 2 roughly corresponds to $\epsilon \sim 0.37$ (see the text).

For WFC3-like data (Figure 15), the inclusion of correlation strength is supported with strong evidence only in the strong correlation ($\epsilon = 0.5$) cases and is otherwise not easily ruled out either way. Interestingly, the inclusion of the error multiple is disfavored with substantial or stronger evidence not only when the error bars are correct, but also even in the case of underestimated error bars. This shows that, for WFC3-like data, the error multiple parameter is not warranted in general.

For NIRSPEC-like data (Figure 16), the inclusion of the correlation strength can be more robustly judged. Its inclusion is supported with strong to decisive evidence when correlation is present. Conversely, its removal is supported with substantial to strong evidence when it is not present. This indicates that, for JWST-like data, there is the possibility that we can characterize the correlation during the retrieval. The inclusion of the error multiple can be more robustly judged as well. Its removal is supported when the error bars are correct, and its inclusion is favored when the error bars are overestimated. However, when the error bars are underestimated, its inclusion is supported with substantial evidence only when there is no correlation. This behavior matches with the results from the previous section that correlation can be mistaken for underestimated error bars. This shows that the error multiple is generally not effective at indicating accommodating underestimated error bars.

The above results show that while it is difficult to conclusively infer the presence of correlated noise with HST data, it is certainly within possibility that its presence and strength can be measured with JWST data. A few caveats must

be made. Our preliminary test presented here treats the data over the entire NIRSPEC wavelength range as sharing correlation, hence providing an abundant number of points for the correlation strength to be measured; in reality, if its discrete grisms are used, any correlation of instrumental origin will be per each wavelength range. Additionally, as we will discuss in Section 5.1, missing physics in the model acts as a source of systematics, which we do not consider here. Furthermore, we used the same noise model to generate the observation instance as well as to retrieve its parameter. While doing so is obviously a gross simplification, especially considering that numerous sources of correlated noise can operate simultaneously, this provides a reasonable starting point toward using a more complex likelihood function to fit for correlated noise. Additionally, adding hyperparameters to a retrieval further dilutes the noise budget, broadening the retrieved uncertainties of other parameters. Ascertaining what degeneracy this incurs on the estimation of other parameters is left for future work.

5. Discussion

5.1. Model Limitations

A major compounding issue is that, when retrieving on real data, model assumptions and unknowns contribute to and act as systematic errors in addition to the data systematics themselves. In short, bad data are degenerate with bad models. In our study we generated the synthetic observations using the identical forward model as that used in the retrieval in order to minimize

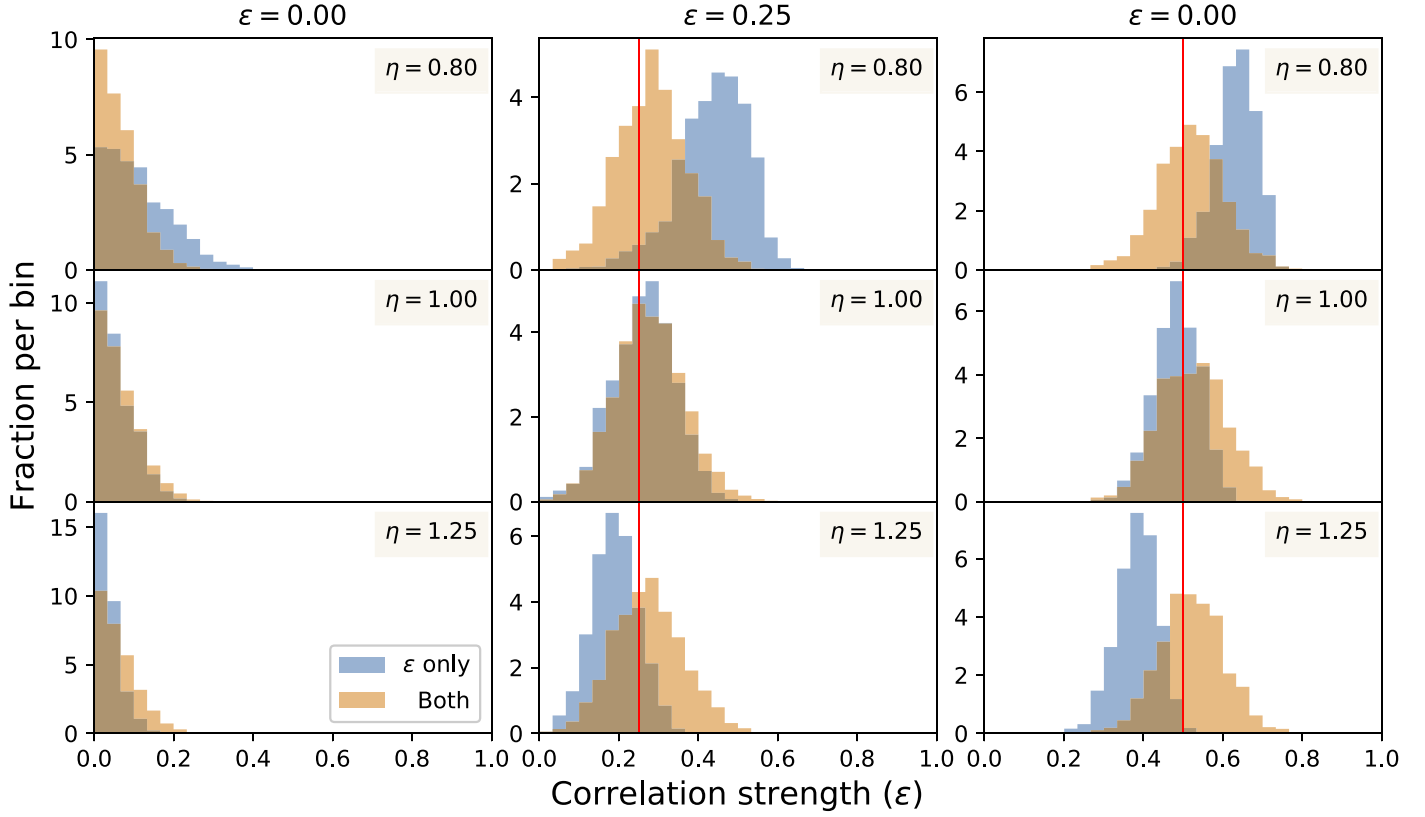


Figure 14. Same as Figure 13, but for NIRSPEC-like wavelength grid. The horizontal scale has been kept the same as in Figure 13 and reflects the full width of the uniform prior used.

Error mult.		$\varepsilon = 0$	$\varepsilon = 0.25$	$\varepsilon = 0.5$
	$\eta = 0.8$	0.7	0.8	0.8
	$\eta = 1.0$	0.6	1.1	1.0
	$\eta = 1.25$	-0.8	0.0	0.0

Correlation str.		$\varepsilon = 0$	$\varepsilon = 0.25$	$\varepsilon = 0.5$
	$\eta = 0.8$	0.4	0.0	-1.2
	$\eta = 1.0$	0.3	0.1	-1.1
	$\eta = 1.25$	0.3	0.0	-1.4

Decisive
Strong
Substantial
Substantial
Strong
Decisive

Inclusion favored -2.0 -1.0 -0.5 B_{10} 0.5 1.0 2.0 Exclusion favored

Figure 15. Bayes factors B_{10} of the retrieval models that exclude error multiple (top) or correlation strength (bottom) for WFC3-like data. The model that includes both hyperparameters is used as the baseline. A positive value indicates that the model better fits the data while spanning a smaller prior volume and hence supports the removal of the parameter. The interpretation of the strength of the evidence is shown in the color bar. The Bayes factors are calculated as the difference in estimated Bayesian log-evidence; the values in the table have a resulting uncertainty of ~ 0.3 .

any model-dependent effects and to isolate the effects of data systematics. In interpreting real data, the fact that our forward models are a simplified incomplete representation of complex atmospheric phenomena will act as a source of systematic error that will remain pervasive, even if the observed data were perfect and free from their own systematics. We therefore

Error mult.		$\varepsilon = 0$	$\varepsilon = 0.25$	$\varepsilon = 0.5$
	$\eta = 0.8$	-0.7	-0.1	0.4
	$\eta = 1.0$	1.3	1.4	1.1
	$\eta = 1.25$	-2.0	-2.0	-1.7

Correlation str.		$\varepsilon = 0$	$\varepsilon = 0.25$	$\varepsilon = 0.5$
	$\eta = 0.8$	0.9	-1.6	-7.8
	$\eta = 1.0$	1.1	-1.5	-7.6
	$\eta = 1.25$	1.1	-1.6	-7.4

Decisive
Strong
Substantial
Substantial
Strong
Decisive

Inclusion favored -2.0 -1.0 -0.5 B_{10} 0.5 1.0 2.0 Exclusion favored

Figure 16. Same as Figure 15, but for the NIRSPEC-like spectrum.

remain open to the possibility that the observed examples of potential systematic noise in the data are in fact due to unaccounted for obscure physics.

For the same reasons discussed above, this will adversely affect high S/N observations in particular, in which the fine (and the not-so-fine) details of the model become discernible. There has recently been a growing body of work that studies the biases incurred by model assumptions and parameterization. To list a few examples for demonstration, MacDonald et al. (2020) performed one-dimensional retrievals on three-dimensional synthetic spectra to show that the retrieval biases the limb temperature to few hundred Kelvins cooler than the

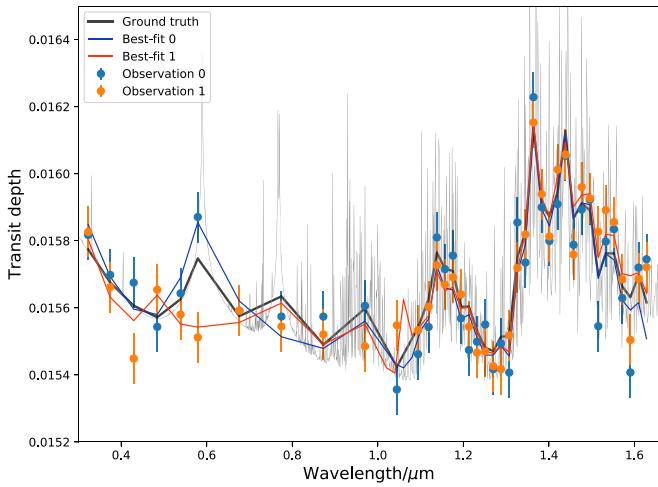


Figure 17. Two observation instances (blue and orange) of the ground-truth spectrum (black) for our baseline hot Jupiter case detailed in Table 1. The best-fit spectra for each observation are also shown.

actual day–night mean temperature. Lacy & Burrows (2020) extended this study to cloudy atmospheres, finding that the presence of aerosols exacerbate the biases induced by three-dimensional effects when not accounted for. Changeat et al. (2019) found that using a vertically constant chemical abundance profile may no longer be sufficient to fully capture signatures of disequilibrium processes in the spectrum. Perhaps most inconspicuously, Barstow et al. (2020) compared and performed cross-retrieval between retrieval codes developed by three groups and found that JWST-quality data is now sensitive to rather rudimentary model unknowns such as the line lists used to generate the opacities and the precision of fundamental constants used.

In our framing of describing biases, these results generally effect shifts in the estimated means of the retrieved parameters. Incorporating our conclusion that correlated noise generally leads to an underestimated error of the retrieved parameters means that biases due to model limitation now strike with a stronger statistical significance. Furthermore, the wavelength-dependent effect of both missing physics and systematics now leave the possibility for degenerate interpretations.

5.2. Instrumental Systematics

To better understand the significance of our results, it would be useful to consider the different sources of how systematics can arise and evaluate their prevalence especially in the context of future space missions such as JWST. While other sources of systematics are possible, such as starspots (Rackham et al. 2018; Bruno et al. 2020), inaccurate orbital parameters, or time-dependent telluric contamination in ground-based observations, the key source of systematics that we will discuss here is instrumental. However, we remind the reader that our formulation of correlated noise can be generalized to any effect that results in wavelength-dependent correlation.

While instrumental systematics are expected to be ubiquitous to some extent, the exact magnitude of their effect in generating wavelength-correlated noise has not been fully understood. Yet a handful of observations exist that hint at the existence of such systematics. Colón et al. (2020) argued that, with current facilities, these systematics are visible at the highest level of precision (~ 15 ppm), inferred from an unusual behavior of the

residuals in the H_2O band. In the spectrum of HD 97658b in Guo et al. (2020), our motivating example in Section 1, we inferred from the inability to fit the data as well as the fact that no obvious physics were missing that there must be some systematics present, even at a lower precision (~ 25 ppm). These examples indicate that some wavelength-correlated noise must be present, unless there is unaccounted for physics in the retrieval model.

There is some reason to surmise that these systematics are more prevalent (or, at least, more noticeable) in the case of bright host stars. A brighter host star allows for a better S/N and higher precision and thereby naturally makes the presence of these systematics more conspicuous compared to a lower S/N data. Additionally, even at the same data quality, a brighter host star requires fewer stackings of observations; for a dimmer host star, by contrast, the number of stackings required to achieve the same S/N naturally averages out any *nonrepeatable* correlated noise. Finally, as alluded to in Section 1, given that instrumental systematics can also behave differently with bright sources, it is not out of the question that there is a separate effect at play here beyond the S/N which may persist through multiple observations in a repeatable fashion.

Comparing the WFC3 spectra of GJ 1214b (Kreidberg et al. 2014) and of HD 97658b (Guo et al. 2020) illustrates this point. While both planets are comparable sub-Neptunes with featureless spectra and have a similar level of precision, the spectrum of the latter displays an unusual upward trend in transit depth in the redder end and other wavelength-correlated residuals throughout the WFC3 bandpass. The relevant difference here may be the host star brightness (9.8 versus 6.2 in the J -band magnitude, respectively). The spectrum of GJ 1214b is the combination of stacking 15 transits, whereas that of HD 97658b has 4. Further, the spectrum of HD 97658b presented in Knutson et al. (2014), which only had the first two visits, shows the most obvious possible example of correlated noise due to systematics.

These types of systematics may be even more pernicious for future high S/N observations from JWST for a few reasons. First, as the noise floor is lower, correlated noise will be relatively more prominent even if it actually manifests at weaker levels. Then one can no longer reliably assume that the observed noise is strictly photon-dominated. This requires an additional step of modeling out now wavelength-dependent systematics during the data reduction, which is necessarily (although perhaps not practically) incomplete. Given that, secondly, the high S/N per transit means that stacking will be unnecessary for most targets. As such, nonrepeating systematics do not get averaged out. Third, we have demonstrated that higher precision leads to biases of stronger significance. This is true even in the absence of systematics in the sense that a retrieval will be more sensitive to the observational instance. Fourthly, we can predict that our understanding of the characteristics of JWST instruments and their appropriate data reduction tools will be only partially correct, at least during the initial few cycles of JWST before practical experience accumulates. Finally, as planets around bright host stars allow for achieving high S/N, they will make attractive targets for JWST observation. However, if the above intuition that bright host stars can exacerbate instrumental systematics is true, it adds another dimension to consider when selecting targets for observation, in addition to the S/N.

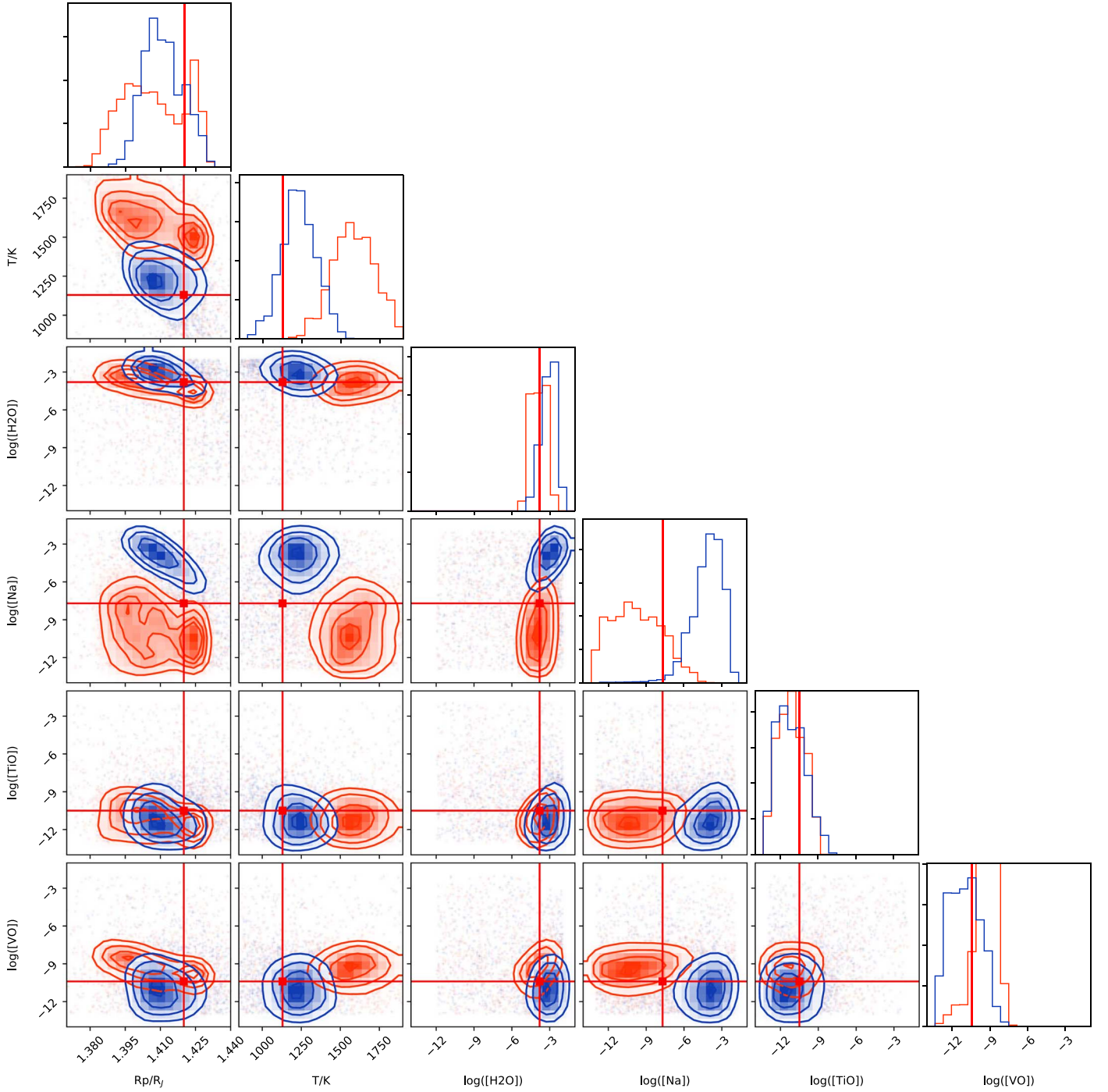


Figure 18. The retrieved posterior from the two input data sets from Figure 17. We only show the relevant parameters. The colors corresponding to each data set are the same as in Figure 17. Black crosshairs indicate the ground-truth input values.

Given that this is the case, it would be worthwhile to put the above heuristic that bright stars bring about correlated noise to a more formal test. This can be accomplished with JWST if the correlation strength can actually be measured, and by marginalizing over the magnitude of the host star to obtain a trend. While this would not comprise a main scientific objective of any program, correlation strength is a parameter we can try to measure for all observations, so this is a test we can perform at no extra cost in observation time.

5.3. Data Outliers and Free Retrieval

PLATON originally supports equilibrium chemistry retrievals only. Using this method the molecular abundance of each species at a given temperature and pressure is set by the metallicity and carbon-to-oxygen ratio. A popular alternative method to constrain the chemistry is to use free retrievals, in which the abundances of each species are allowed to vary independently. This accounts for any nonequilibrium chemistry effects in the atmosphere, brought on by vertical mixing or photochemical interactions. To establish

how wavelength-dependent systematics or outliers can bias certain species, we implement free retrievals in PLATON and perform some basic tests in addition to the suite of retrievals already presented that used equilibrium chemistry models.

PLATON already naively supports inputting custom chemical profiles to its forward models, but only accepts equilibrium chemistry parameters during retrievals. We extend its capability by allowing it to accept custom chemical abundances during retrievals as well. As such, all other details regarding how the spectrum is calculated during the retrieval remain exactly the same as the original implementation in PLATON. We assume that each species has a vertically fixed mixing ratio. This is a reasonable approximation over the pressure range probed by transit spectroscopy at the present data quality, and most current one-dimensional free retrieval codes parameterize the composition using this assumption (Caldas et al. 2019). JWST-quality data may merit a more complex prescription, such as a two-part vertical abundance profile (Changeat et al. 2019), but for now we do not consider these possibilities. Additionally, for these tests we remove the simulated Spitzer data points from our synthetic spectra to limit the wavelength range, thereby reducing the choice of necessary species to be included. We assume an H/He-dominated background atmosphere. We include H₂O, Na, K, TiO, and VO as they are the primary detectable species in the remaining wavelength and temperature range; we are mostly interested in how differently species with broad absorption features (e.g., H₂O) versus species with narrow ones (e.g., Na) can be biased.

We show a free retrieval on two observational instances of the same baseline hot Jupiter as described in Section 2, where we did not add correlated noise or any instrumental offsets. Figure 17 shows the two random input data realizations and the best-fit spectra, and Figure 18 shows the retrieved posteriors.

The point of interest here is how the retrieved abundances compare for species with broad spectral features (H₂O, 1–2 μ m) and with a narrow feature (Na, 0.58 μ m). As one may expect, for both of the two random spectra, the retrieved posterior distribution for water shows a tighter constraint than that for Na, which is dictated almost entirely by one data point. Consequently, between the two posteriors as well, the retrieved distributions for Na show little overlap, ruling out each other by $\geq 2\sigma$. These are simply two randomly drawn samples, but it demonstrates the point that measurements of water abundance or metallicity are more robust compared to that of say Na or K abundances, because of the impact the former parameters have across a broader wavelength range.

6. Summary and Future Work

Atmospheric retrieval provides a robust framework to interface theory and observations and is a key tool to furthering our understanding of exoplanets. One major outstanding issue is disentangling the effects of systematic biases that may be in operation, and in response there is a growing body of work in the literature that investigates the consequences of biases that arise from forward model assumptions.

This paper instead presents an assessment of biases that arise from systematic noise in data, while remaining agnostic as to the source of such systematics. We stress that, although our implementation of correlated noise (using a Gaussian process) is just one mathematical option, the general results remain robust. We find that the presence of correlated noise can mislead us in various ways. We are more likely on average to

obtain a better goodness of fit, but obtain worse retrieval accuracy overall. This is due to both the parameter mean being biased and the retrieved error being underestimated. Specifically, we observe that correlated noise can bias the retrieved aerosol properties, mimicking non-Rayleigh slopes or misrepresenting the location of a cloud deck. Additionally, we find that offsets between data sets can be correctly retrieved and are not degenerate with the retrieved chemistry when equilibrium chemistry is assumed, so long as the forward model is an accurate depiction of the atmosphere. We also find that while correlated noise cannot be characterized during retrieval for HST data, there is potential (and perhaps necessity) for JWST data, even though our tests reflect optimistic conditions. Additionally, we validate the intuition that retrievals are sensitive to individual noise instances, and, especially in the context of free retrievals, that statistical outliers can have significant effects on the retrieved chemistry.

We thank Jacob Bean, Drake Deming, and Kevin Stevenson for useful conversations about data systematics in exoplanet spectra and Cole Miller for discussions on the finer points of Bayesian statistics. J.I. and E.M.-R.K. acknowledge support from the National Science Foundation under CAREER grant No. 1931736 and from the Research Corporation for Science Advancement through their Cottrell Scholar program. Simulations were performed on the YORP cluster administered by the Center for Theory and Computation, part of the Department of Astronomy at the University of Maryland.

ORCID iDs

Jegou Ih  <https://orcid.org/0000-0003-2775-653X>
Eliza M.-R. Kempton  <https://orcid.org/0000-0002-1337-9051>

References

- Andrae, R., Schulze-Hartung, T., & Melchior, P. 2010, arXiv:1012.3754
- Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, *ARA&A*, 47, 481
- Barstow, J. K., Changeat, Q., Garland, R., et al. 2020, *MNRAS*, 493, 4884
- Barstow, J. K., & Heng, K. 2020, *SSRv*, 216, 82
- Benneke, B., Knutson, H. A., Lothringer, J., et al. 2019, *NatAs*, 3, 813
- Bruno, G., Lewis, N. K., Alam, M. K., et al. 2020, *MNRAS*, 491, 5361
- Caldas, A., Leconte, J., Selsis, F., et al. 2019, *A&A*, 623, A161
- Changeat, Q., Edwards, B., Waldmann, I. P., & Tinetti, G. 2019, *ApJ*, 886, 39
- Cobb, A. D., Himes, M. D., Soboczenski, F., et al. 2019, *AJ*, 158, 33
- Colón, K. D., Kreidberg, L., Welbanks, L., et al. 2020, *AJ*, 160, 280
- Deming, D., & Seager, S. 2017, *JGR*, 122, 53
- Deming, D., Sada, P. V., Jackson, B., et al. 2011, *ApJ*, 740, 33
- Deming, D., Wilkins, A., McCullough, P., et al. 2013, *ApJ*, 774, 95
- Dittmann, J. A., Close, L. M., Green, E. M., Scuderi, L. J., & Males, J. R. 2009, *ApJ*, 699, L48
- Evans, T. M., Aigrain, S., Gibson, N., et al. 2015, *MNRAS*, 451, 680
- Feroz, F., & Hobson, M. P. 2008, *MNRAS*, 384, 449
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
- Fortney, J. J., Visscher, C., Marley, M. S., et al. 2020, *AJ*, 160, 288
- Garhart, E., Deming, D., Mandell, A., et al. 2020, *AJ*, 159, 137
- Gennaro, M. E. A. 2018, WFC3 Data Handbook, Version 4.0 (Baltimore, MD: STScI)
- Guo, X., Crossfield, I. J. M., Dragomir, D., et al. 2020, *AJ*, 159, 239
- Hou Yip, K., Changeat, Q., Edwards, B., et al. 2020, *AJ*, 161, 4
- Knutson, H. A., Charbonneau, D., Noyes, R. W., Brown, T. M., & Gilliland, R. L. 2007, *ApJ*, 655, 564
- Knutson, H. A., Dragomir, D., Kreidberg, L., et al. 2014, *ApJ*, 794, 155
- Kreidberg, L., Bean, J. L., Désert, J.-M., et al. 2014, *Natur*, 505, 69
- Lacy, B. I., & Burrows, A. S. 2020, *ApJ*, 905, 131
- Line, M. R., Vasisth, G., Chen, P., Angerhausen, D., & Yung, Y. L. 2011, *ApJ*, 738, 32

- Line, M. R., Wolf, A. S., Zhang, X., et al. 2013, [ApJ](#), **775**, 137
- Lothringer, J. D., Benneke, B., Crossfield, I. J. M., et al. 2018, [AJ](#), **155**, 66
- MacDonald, R. J., Goyal, J. M., & Lewis, N. K. 2020, [ApJL](#), **893**, L43
- Madhusudhan, N. 2018, in *Handbook of Exoplanets*, ed. H. J. Deeg & J. A. Belmonte, 104 (Cham: Springer), 104
- Marley, M. S., Ackerman, A. S., Cuzzi, J. N., & Kitzmann, D. 2014, in *Comparative Climatology of Terrestrial Planets*, ed. S. J. Mackwell et al. (Tucson, AZ: Univ. of Arizona Press), 367
- May, E. M., Gardner, T., Rauscher, E., & Monnier, J. D. 2020, [AJ](#), **159**, 7
- Miller-Ricci Kempton, E., Zahnle, K., & Fortney, J. J. 2012, [ApJ](#), **745**, 3
- Pinhas, A., Rackham, B. V., Madhusudhan, N., & Apai, D. 2018, [MNRAS](#), **480**, 5314
- Rackham, B. V., Apai, D., & Giampapa, M. S. 2018, [ApJ](#), **853**, 122
- Rodgers, C. D. 2000, *Inverse Methods for Atmospheric Sounding* (Singapore: World Scientific)
- Sing, D. K., Fortney, J. J., Nikolov, N., et al. 2016, [Natur](#), **529**, 59
- Sivia, D., & Skilling, J. 1996, *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford Univ. Press)
- Stevenson, K. B., & Fowler, J. 2019, Instrument Science Report WFC3 2019-12, Space Telescope Science Institute
- Taylor, J., Parmentier, V., Irwin, P. G. J., et al. 2020, [MNRAS](#), **493**, 4342
- Tsiaras, A., Waldmann, I. P., Rocchetto, M., et al. 2016, [ApJ](#), **832**, 202
- Waldmann, I. P., Tinetti, G., Rocchetto, M., et al. 2015, [ApJ](#), **802**, 107
- Zhang, M., Chachan, Y., Kempton, E. M.-R., & Knutson, H. A. 2019, [PASP](#), **131**, 034501