Correcting Exoplanet Transmission Spectra for Stellar Activity with an Optimized Retrieval Framework

The Astrophysical Journal, 960:107 (26pp), 2024 January 10 https://doi.org/10.3847/1538-4357/ad0369

Correcting Exoplanet Transmission Spectra for Stellar Activity with an Optimized Retrieval Framework

Alexandra Thompson[1], Alfredo Biagini[2,3], Gianluca Cracchiolo[2,3], Antonino Petralia[3], Quentin Changeat[1,4], Arianna Saba[1], Giuseppe Morello[5,6,7], Mario Morvan[1], Giuseppina Micela[3,7], and Giovanna Tinetti[1]
[1] Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT London, UK
[2] University of Palermo, Department of Physics and Chemistry 'Emilio Segrè' Via Archirafi, 36, I-90123 Palermo, Italy
[3] INAF-Osservatorio Astronomico di Palermo, Piazza del Parlamento, 1, I-90134 Palermo, Italy
[4] European Space Agency (ESA), ESA Office, Space Telescope Science Institute (STScI), 3700 San Martin Drive, Baltimore, MD 21218, USA
[5] Instituto de Astrofísica de Canarias (IAC), E-38205 La Laguna, Tenerife, Spain
[6] Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38206, La Laguna, Tenerife, Spain
[7] Department of Space, Earth and Environment, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden
Received 2023 June 2; revised 2023 September 27; accepted 2023 October 12; published 2024 January 4

Abstract

The chromatic contamination that arises from photospheric heterogeneities, e.g., spots and faculae on the host star presents a significant noise source for exoplanet transmission spectra. If this contamination is not corrected for, it can introduce substantial bias in our analysis of the planetary atmosphere. We utilize two stellar models of differing complexity, `StARPA` (Stellar Activity Removal for Planetary Atmospheres) and `ASteRA` (Active Stellar Retrieval Algorithm), to explore the biases introduced by stellar contamination in retrieval under differing degrees of stellar activity. We use the retrieval framework TauREx3 and a grid of 27 synthetic, spot-contaminated transmission spectra to investigate potential biases and to determine how complex our stellar models must be in order to accurately extract the planetary parameters from transmission spectra. The input observation is generated using the more complex model (`StARPA`), in which the spot latitude is an additional, fixable parameter. This observation is then fed into a combined stellar-planetary retrieval, which contains a simplified stellar model (`ASteRA`). Our results confirm that the inclusion of stellar activity parameters in retrieval minimizes bias under all activity regimes considered. `ASteRA` performs very well under low-to-moderate activity conditions, retrieving the planetary parameters with a high degree of accuracy. For the most active cases, characterized by larger, higher-temperature contrast spots, some minor residual bias remains due to `ASteRA` neglecting the interplay between the spot and the limb-darkening effect. As a result of this, we find larger errors in retrieved planetary parameters for central spots (0°) and those found close to the limb (60°) than those at intermediate latitudes (30°).

Unified Astronomy Thesaurus concepts: Stellar activity (1580); Exoplanet atmospheres (487); Exoplanets (498); Transmission spectroscopy (2133)

1. Introduction

With over 5000 confirmed exoplanet detections, and this number rapidly increasing with the contribution of ground-based surveys, e.g., HARPS (Mayor et al. 2003) and space-based missions such as TESS (Ricker et al. 2014), we are entering a period of unprecedented potential for exoplanet characterization. Present exoplanet observations and analyses are laying the foundations for the large-scale population studies that will be conducted with next-generation space observatories such as JWST (Bean et al. 2018), and in less than a decade, Ariel (Tinetti et al. 2021). Planets from multiple, distinct regions of the known parameter space have already been analyzed in detail with transmission (e.g., Charbonneau et al. 2002; Tinetti et al. 2007; Sing et al. 2016; Tsiaras et al. 2018, 2019; Hoeijmakers et al. 2019; Pinhas et al. 2019; Anisman et al. 2020; Pluriel et al. 2020; Skaf et al. 2020; Edwards et al. 2023; Gressier et al. 2022; Saba et al. 2022; Rustamkulov et al. 2023), emission (e.g., Swain et al. 2008; Crouzet et al. 2014; Evans et al. 2017; Mikal-Evans et al. 2020; Changeat et al. 2022), and phase curve spectroscopy (e.g., Knutson et al. 2012; Stevenson et al. 2014; Arcangeli et al. 2019; Feng et al. 2020; Irwin et al. 2020; von Essen et al. 2020; Changeat 2022; Dang et al. 2022). All of these techniques come with the caveat that the planet and star are observed as an unresolved source, with the host star providing the light source that makes these methods possible. As such, disentangling the stellar signals from the planetary ones is a challenging, but essential part of any exoplanet characterization pipeline. The objective of this paper is to outline a simple, scalable method of disentangling these signals that can be implemented seamlessly in retrievals. The primary aim is to accurately retrieve the planetary parameters in the presence of stellar activity. We investigate what biases the simplified model assumptions could produce as a result of missing physics and how limiting these biases could be. As a secondary aim, we also explore what useful information about the host star can be extracted from a combined stellar-planetary retrieval.

Our current understanding of exoplanet host stars indicates that a substantial fraction of them will display moderate-to-high levels of activity. As such, stellar contamination will likely be one of the most dominant noise sources in exoplanetary observations. Evidence for this is shown in the form of activity indicators, e.g., Ca H-K lines, S-index, etc. (Gomes da Silva et al. 2011; Cauley et al. 2018; Klein et al. 2022), variability amplitudes from long-term photometric monitoring, e.g., with

Kepler (Ciardi et al. 2011; McQuillan et al. 2014) and the many spot and plage-crossing events that have been detected in light curves of transiting exoplanets (e.g., Pont et al. 2013; Oshagh et al. 2014; Morris et al. 2017; Espinoza et al. 2019). With higher-resolution observations rapidly becoming available to us it will be essential to account for this activity, ideally in a homogeneous way, in order to characterize exoplanet atmospheres with the high precision we are aiming for and subsequently to conduct larger, comparative population studies. Observations covering larger wavelength ranges will also be highly beneficial as they allow for an easier identification and subsequent correction for stellar contamination.

Some trends in activity have already been well documented in the literature, with higher levels of activity typically observed for later-type K-M dwarfs (Goulding et al. 2012; Jackson & Jeffries 2013; McQuillan et al. 2014) as stars transition toward becoming fully convective (Reiners & Basri 2010) and also in the case of fast rotating, young stars or young stellar objects (YSOs; e.g., Gully-Santiago et al. 2017; Järvinen et al. 2018; Morris 2020; Klein et al. 2022). These scenarios are particularly important as much of the pioneering exoplanetary science is now focusing on planets orbiting these stars. These smaller, later-type stars are frequently observed to host small planets (Dressing & Charbonneau 2015), which are crucial for pushing the limits of our current characterization techniques, and ultimately for answering questions surrounding potential habitability. Contamination due to stellar activity has the potential to negatively affect observations of all types of planets, albeit in different ways. For larger planets with H-dominated, primary atmospheres, the bias introduced by not accounting for contamination could potentially result in inaccurate retrieved atmospheric parameters such as chemical abundances and temperature-pressure profiles (see, e.g., Saba et al. 2022). These inaccuracies may subsequently be propagated into discussions surrounding other key research areas, e.g., elemental ratios (Öberg et al. 2011; Pacetti et al. 2022) and disequilibrium chemistry (Venot et al. 2020) leading to an incorrect interpretation of the planet as a whole. For small planets, particularly those possessing secondary atmospheres, atmospheric absorption features will generally tend to be weaker, making them much more susceptible to being obscured or altered by stellar contamination (e.g., Ballerini et al. 2012; Rackham et al. 2018; Zhang et al. 2018). In recent years, the number of observations of exoplanets orbiting YSOs/young main-sequence stars, e.g., AU Mic (Szabó et al. 2021; Klein et al. 2022), V1298 Tau (Feinstein et al. 2022; Maggio et al. 2022), and K2-33b (Thao et al. 2023), has also increased as, despite being strongly affected by stellar activity, studying these systems provides us with a unique opportunity to begin to fill in gaps in our understanding of the stages of planetary formation and evolution (Raymond & Morbidelli 2022).

In transmission spectroscopy, an active star is capable of contaminating the observed spectrum in multiple ways by causing the transit chord to differ from the disk-integrated stellar spectrum. The most efficient methods of modeling and correcting for this contamination have been and are still being explored extensively (e.g., Czesla et al. 2009; Garcia et al. 2010; Silva-Valio et al. 2010; Sing et al. 2011; Ballerini et al. 2012; Oshagh et al. 2014; Micela 2015; Herrero et al. 2016; Newton et al. 2016; Zellem et al. 2017; Rackham et al. 2018; Bixel et al. 2019; Cracchiolo et al. 2021a, 2021b). At the spectral intervals observed for exoplanet characterization, typically the optical and near-infrared (NIR) regimes, we are observing the stellar photosphere. On the photosphere, magnetic activity manifests in the form of heterogeneities such as cooler spots and hotter faculae (Solanki 2003; Berdyugina 2005). If present, these features are the dominant sources of stellar contamination in this wavelength range. The strength of this contamination and its overall effect on the observed spectrum differs depending on what type of active regions are present and whether or not they are occulted by the transiting planet. Occulted active regions are arguably easier to identify due to the characteristic bumps they introduce in the light curves. These bumps, while not always easily corrected for, can often be masked/removed from the light curve. Unocculted features are more insidious as they affect the transit depth of the entire light curve, without imparting any obvious identifying feature on it. In addition to this, unocculted features could potentially be far more common, as the planet only occults a small fraction of the stellar disk as it transits.

Unocculted spots, the focus of this study, reduce the average flux that originates from the regions of the star not crossed by the planet. Their presence introduces a wavelength-dependent signal that deepens the transit light curve, resulting in overestimates of the planetary radius, particularly in the optical regime. Stellar contamination is highly chromatic, with its effects appearing substantially stronger and more evident at shorter wavelengths (e.g., Ballerini et al. 2012; Rackham et al. 2018). The main concern when contamination is present but not corrected for is that it may introduce biases in the retrieved planetary parameters. It is capable of both obscuring absorption features in the case of unocculted faculae, or mimicking/strengthening them in the case of unocculted spots. An additional complication arises in that stellar contamination is temporally variable. Modulation occurs predominantly on the timescale of stellar rotation but also through spot evolution and longer-timescale magnetic cycles (Ciardi et al. 2011; Bradshaw & Hartigan 2014; McQuillan et al. 2014; Zhang et al. 2018). This means that each observation of each exoplanetary system should ideally be corrected individually before they can accurately be combined or analyzed simultaneously. This limitation poses a problem for small planets in particular, as we will need to stack observations from multiple visits to obtain a high enough signal-to-noise ratio (S/N) for a successful retrieval analysis.

For the above reasons, stellar activity is one of the most pressing challenges to the accurate characterization of exoplanetary atmospheres at present. Modeling the star as a more complex astrophysical body, rather than as a homogeneous light source, is essential in tackling this issue. Despite this, increases in model complexity are not always beneficial and care needs to be taken when introducing additional parameters in retrievals as this can have detrimental effects. Using models with a higher dimensionality than is necessary may increase the risk of overfitting the data or injecting a bias through the model choice. Alternatively, if many of the parameters are degenerate, the retrieval may not be able to converge on a solution at all. More complex models will also intrinsically come with an associated computational cost. The aim of this paper is to find the middle ground by determining a stellar model that encompasses all of the essential physics of stellar activity required to remove any potential biases, but without overcomplicating the model to the detriment of retrieval reliability

or computing time. To achieve this we utilize two approaches to modeling stellar activity in the form of a single, unocculted spot. The predominant difference between them is the consideration of the limb-darkening effect, or the lack thereof. StARPA (Stellar Activity Removal for Planetary Atmospheres), the more complex of the two models, encompasses the interplay of the presence of spots and the limb-darkening effect by fixing the spot position on the stellar disk. The StARPA model is then used as the forward model in a benchmarking retrieval exercise that uses ASteRA (Active Stellar Retrieval Algorithm), the simpler of the two models, in which this interplay between spot contamination and limb darkening is neglected and all parameters describing the spot and the planetary atmosphere are fitted simultaneously. Stellar activity has been considered and fit for in several previous retrieval studies (Pinhas et al. 2018; Bixel et al. 2019; Espinoza et al. 2019; Edwards et al. 2021). In order to obtain accurate planetary parameters and maximize the information content from our retrievals we need to have a good understanding of what the potential biases from stellar activity are. These biases can originate both from neglecting stellar contamination or could be introduced or incompletely removed by our chosen correction process. Retrieval biases due to not correcting for stellar activity have previously been systematically explored in Iyer & Line (2020). This work presents the first investigation into the potential residual biases that are left behind as a result of neglecting the limb-darkening effect in the correction process. A similar investigation was conducted in Cracchiolo et al. (2021b), albeit on a smaller scale. In Section 2.2 we introduce our grid of 27 spot-contaminated stellar disks from which we generate the contaminated transmission spectra. This grid is produced using the StARPA model where the activity is characterized by a single-unocculted starspot. This spot is parameterized by its temperature contrast with respect to the quiet photosphere, its radius, and the latitude of the spot's center. The StARPA model is described in detail in Section 2.3 followed by a description of the simplified ASteRA model that is used in retrievals in Section 2.4. We conduct retrievals on the grid of contaminated spectra to investigate under which conditions ASteRA is capable of accurately retrieving the planetary, and to a lesser extent, the stellar parameters, in the presence of varying degrees of stellar contamination. The results of these retrievals are given in Section 3. Detailed analysis of these results alongside initial investigations into more complex, realistic cases involving multiple spots and spots that are separated into an umbra and penumbra are given in Section 4. In this section, we also discuss when using a simplified stellar model, such as ASteRA, is valid at first order and under what activity conditions the assumptions within it begin to break down. Our final, concluding remarks are given in Section 5.

2. Methodology

This section introduces the overall experimental framework (Section 2.2) and the two modeling approaches utilized throughout the work (Sections 2.3 and 2.4) to explore the effect of model complexity on the accuracy of the retrieved planetary, and to a lesser extent, stellar parameters. The main objective of this work is to develop a stellar activity correction method that is both computationally fast and capable of retrieving planetary parameters accurately for host stars displaying different levels of activity. It is of course desirable to also be able to accurately retrieve the stellar parameters; however, we prioritize the planetary parameters as these are fundamental for the large-scale characterization of exoplanet populations intended to be carried out with ongoing and future missions (e.g., JWST and Ariel). We aim to determine under which activity conditions using a simplified model is sufficient, and if/where it is insufficient, investigate any potential biases it may introduce.

These questions surrounding model complexity are essential questions to answer because as our underlying physical models become more realistic, their dimensionality increases drastically, which will likely require an increase in computation time, and in the worst scenario can be detrimental to retrieval accuracy. On the contrary, by using an oversimplified model we risk neglecting underlying physical processes that are necessary in order to fully and accurately interpret our observations. As such, we are increasingly facing a compromise between complexity and the computing cost such models require, although state-of-the-art machine-learning methods could potentially mitigate this (e.g., Yip et al. 2022).

To explore the ability of the ASteRA model to accurately retrieve the planetary parameters, we generate a grid of 27 spot-contaminated transmission spectra as forward models. These are produced using the StARPA model (Section 2.3) and are then used as input observations in retrievals where the stellar contamination is accounted for in a combined stellar-planetary retrieval, using TauREx3 and the ASteRA plugin described in Section 2.4. We also introduce a case with an uncontaminated, quiet star, termed Case 0, which acts as a baseline for the subsequent retrieval analysis. Each contaminated stellar disk in our grid is characterized by a single, unocculted spot, which is itself parameterized by its temperature contrast ($\Delta T_{\rm spot}$) i.e., how much cooler than the quiescent photosphere the spot is, the spot radius ($R_{\rm spot}$) normalized to the stellar radius and the latitude of the spot center ($\phi_{\rm spot}$) from which we can probe the effect of limb darkening. These 27 cases are discussed in greater detail in Section 2.2.

2.1. Simulating the Uncontaminated Planetary Transmission Spectrum for a Synthetic Star–Planet System

To determine how extreme the contamination effects are for each spot case considered and explore how well these effects can be mitigated, we must first produce a synthetic star–planet system, for which the stellar and planetary parameters are known a priori. We consider a synthetic K dwarf host star characterized by the following parameters ($T_{\rm eff}$ = 4750 K, $R_\star$ = 0.8 $R_\odot$, $M_\star$ = 0.8 $M_\odot$) and displaying activity in the form of a single-unocculted spot, which we model using the methodology described in Section 2.3. This is the same methodology outlined in Cracchiolo et al. (2021b) with some subsequent, minor improvements to its efficiency. A K dwarf was chosen as stars of later spectral types are typically more likely to be active (Berdyugina 2005; Ciardi et al. 2011; Hartman et al. 2011; McQuillan et al. 2014; Rackham et al. 2018, 2019). We decided against using an M dwarf host for this preliminary study as this is the region of the main sequence in which stars transition to a fully convective regime (Reiners & Basri 2010), as such, stellar activity may manifest differently on these stars. We choose to focus on single-spot cases for this initial benchmarking study for several reasons. First, this represents the simplest spot-contaminated disk model that can be considered, which makes it invaluable as a baseline for future

investigations. Encompassed with this is that it reduces the dimensionality required for the input model. Having multiple spots present requires the location of each of them on the stellar disk to be defined in order to accurately compute the interplay between their properties and the limb-darkening effect. Second, a single, larger spot also allows us to probe the extremes of this interplay in a way that multiple spots will not as, for a given filling factor, the active region is concentrated at a single latitude rather than dispersed over multiple locations on the stellar disk. Despite this, the extension to multiple-spot cases is comparatively straightforward as described in Cracchiolo et al. (2021b). The results of several preliminary multiple-spot cases are presented in Section 4.5 for completeness.

The transiting planet is a temperate sub-Neptune ($R_P = 3\ R_\oplus$ (0.273 $R_{\rm Jup}$), $M_P = 5\ M_\oplus$ (0.0157 $M_{\rm Jup}$), and $T_P = 400$ K) with a primary atmosphere containing water ($\log(H_2O) = -3$) and H and He present as fill gases with a ratio of 0.172. Rayleigh scattering and collision-induced absorption (CIA) are also included and introduce wavelength-dependent contributions to the opacity. A smaller planet with a primary atmosphere is used as its scale height should result in detectable absorption features but the smaller S/N of such features means that they are more susceptible to being masked by stellar contamination. As such, the accuracy of our correction method, and any biases that may be inadvertently introduced, are comparatively far more important here than when considering an atmosphere with a much higher S/N, for example, that of a hot Jupiter. The orbital inclination is set to 88° so that the effects of a central-unocculted spot ($\phi_{\rm spot} = 0°$) can also be explored.

Using the stellar and planetary parameters described above and the retrieval code `TauREx3` (Al-Refaie et al. 2021, 2022), we produce a forward model that is equivalent to the idealized, uncontaminated transmission spectrum that would be observed in the presence of a completely homogeneous host star. This synthetic spectrum has a wavelength coverage of 0.5–9.5 μm and a resolution of 200. This wavelength range has been chosen as it is similar to the regions that are/will be covered by the JWST and Ariel instruments but with greater coverage and resolution in the optical regime where the effects of stellar contamination are strongest. Error bars of 10 ppm are introduced for all data points regardless of wavelength. This noise level was chosen as 10 ppm is broadly consistent with the most precise observations obtained with the Hubble Space Telescope (HST) instruments (e.g., Edwards et al. 2023), albeit for larger planets further into the IR. We acknowledge that obtaining this level of precision, in reality, would be extremely challenging and heavily reliant on an accurate characterization of the host star to mitigate the astrophysical noise. We rationalize our choice of using small error bars as our goal at this stage is to be limited by the model assumptions/limitations and not the instrument performance, which will be a focus of future work. Testing our correction methods on an idealized case first allows us to quantify how effective they are and ensure that they do not introduce any unknown, intrinsic bias before using them with real observations.

2.2. The Experimental Framework: 27 Spot-contaminated Cases

Throughout this study, we consider 27 spot-contaminated cases (and an uncontaminated case as a reference frame) for the star–planet system outlined in the previous section. From this grid, we investigate the potential biases that an unocculted spot may introduce and how these may vary as a function of the three spot parameters considered ($\Delta T_{\rm spot}$, $R_{\rm spot}$, and $\phi_{\rm spot}$). The contaminated stellar disks are produced with `StARPA` (Section 2.3) and the following values are considered for each spot parameter: $\Delta T_{\rm spot} = [250, 500, 1000]$ K, $R_{\rm spot} = [0.1, 0.2, 0.4]\ R_\star$, and $\phi_{\rm spot} = [0, 30, 60]°$. This results in 27 unique parameter combinations in which only one spot parameter is varied at a time. The full set of spot parameters used in each case are given in Table 1. Visual representations of each case are shown in Figure 1.

Table 1
An Outline of the 27 Single-spot Scenarios Considered in This Study

Case No.	$\Delta T_{\rm spot}$ (K)	$R_{\rm spot}$ (R_*)	$\phi_{\rm spot}$ (°)	$F_{\rm spot}$ (%)
0	N/A	N/A	N/A	N/A
1	1000	0.1	0	1
2	1000	0.1	30	0.867
3	1000	0.1	60	0.498
4	1000	0.2	0	4
5	1000	0.2	30	3.462
6	1000	0.2	60	2.001
7	1000	0.4	0	16
8	1000	0.4	30	13.858
9	1000	0.4	60	7.997
10	500	0.1	0	1
11	500	0.1	30	0.867
12	500	0.1	60	0.498
13	500	0.2	0	4
14	500	0.2	30	3.462
15	500	0.2	60	2.001
16	500	0.4	0	16
17	500	0.4	30	13.858
18	500	0.4	60	7.997
19	250	0.1	0	1
20	250	0.1	30	0.867
21	250	0.1	60	0.498
22	250	0.2	0	4
23	250	0.2	30	3.462
24	250	0.2	60	2.001
25	250	0.4	0	16
26	250	0.4	30	13.858
27	250	0.4	60	7.997

Note. Each case is characterized by a unique combination of three spot parameters: the temperature contrast with respect to the photosphere ($\Delta T_{\rm spot}$), the spot radius normalized to the stellar radius ($R_{\rm spot}$), and the latitude of the spot center ($\phi_{\rm spot}$). The spot-filling factor ($F_{\rm spot}$) is calculated assuming an elliptical projection onto the surface when the spot center is at nonzero latitudes. An uncontaminated case is also considered to act as a control case (Case 0).

2.3. The Input Stellar Model: `StARPA`

From the uncontaminated planetary transit spectrum produced with `TauREx`, the grid of 27 spot-contaminated spectra described above is constructed using the Cracchiolo et al. (2021b) model with some subsequent improvements made to the implementation. The key point of this methodology with respect to this study is that, through accounting for the limb-darkening effect and defining the exact position of the spot on the stellar disk, any interactions between contamination and limb darkening can be explored. As in Cracchiolo et al. (2021b), the spots are modeled as being circular but with an elliptical 2D projection on the visible stellar disk at nonzero latitudes. The out-of-transit stellar flux from the spot-

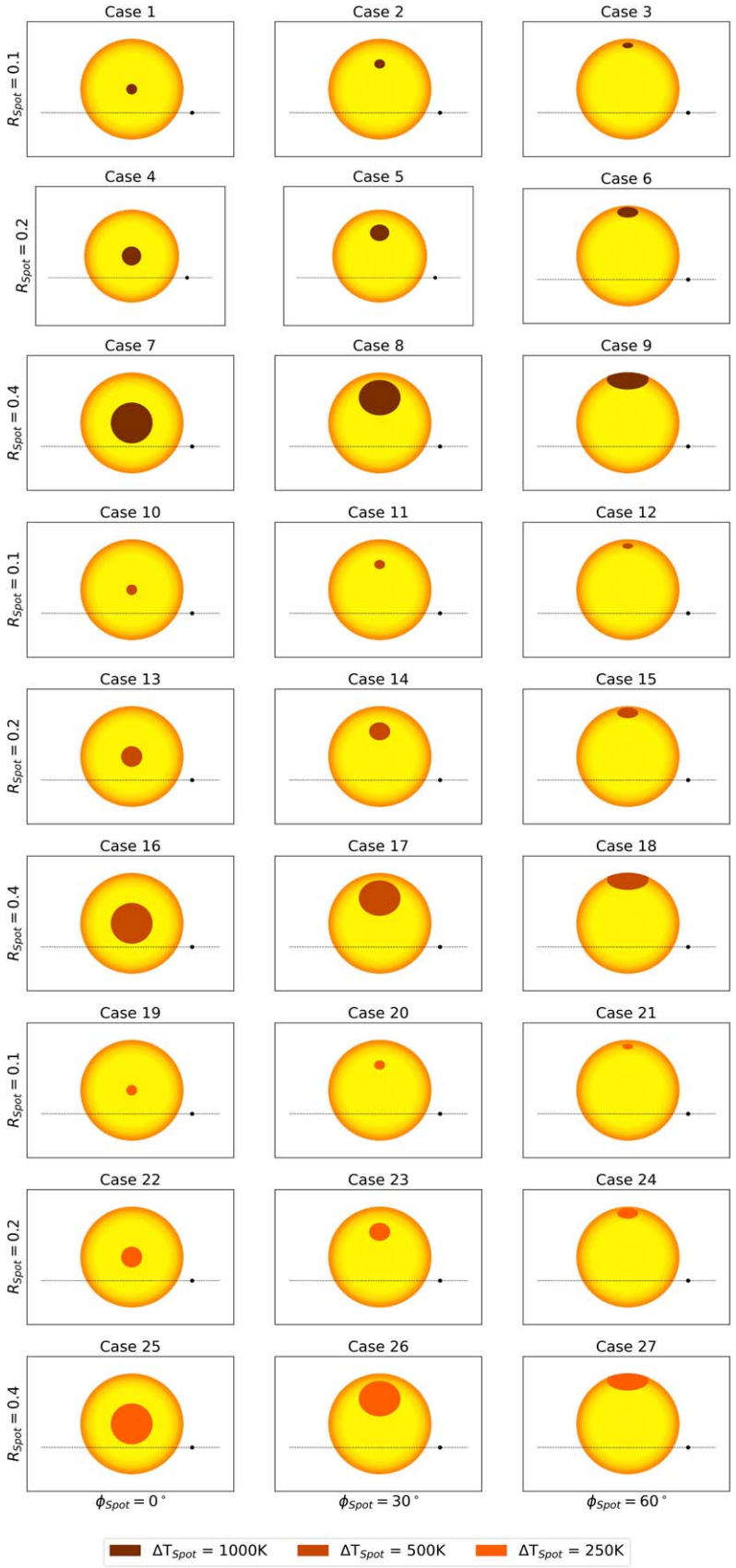

Figure 1. Visual representations of the 27 spotted star cases investigated in this study. The spot color corresponds to its temperature contrast with respect to the quiescent photosphere, which has a temperature of $T_{\rm phot} = 4750$ K.

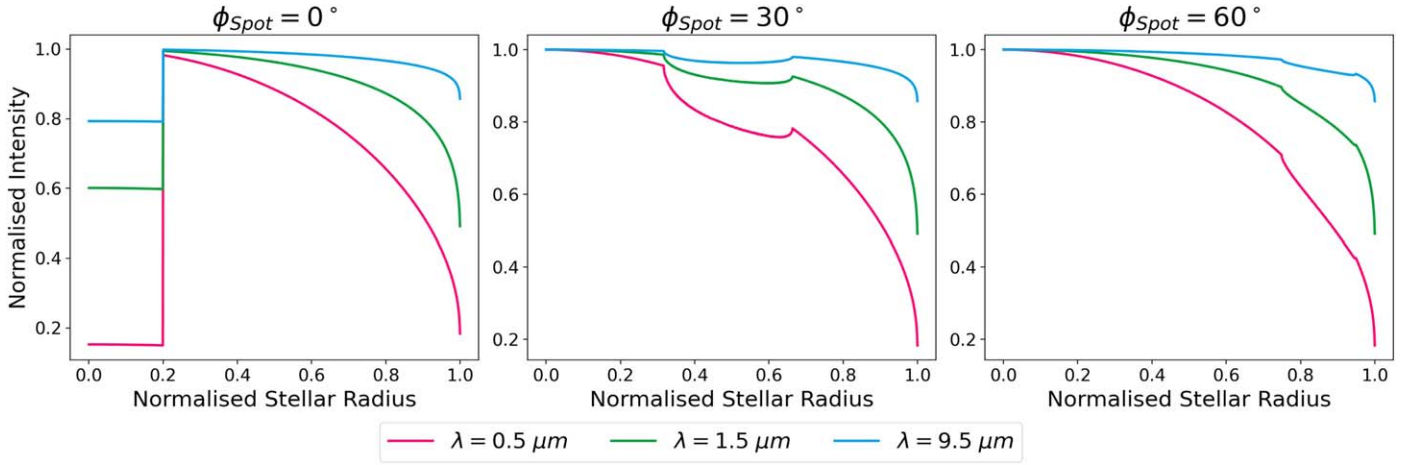

Figure 2. Intensity profiles, normalized to the flux emitted from the disk center of a homogeneous star with $T_{\rm eff} = 4750$ K, for a spot-contaminated star of the same temperature possessing a $0.2R_*$, $\Delta T_{\rm spot} = 1000$ K spot located at latitudes of 0° (left), 30° (center), and 60° (right) and viewed at wavelengths of 0.5 μm (pink), 1.5 μm (green), and 9.5 μm (blue), respectively.

contaminated star is computed using quadrature integration; the star is divided into 1000 equally spaced annuli and the fractions of each of these annuli covered by the spot are calculated. This enables us to compute both the absolute and normalized intensity profiles (Figure 2).

To create the contaminated stellar disks we model the contribution of the quiet photosphere and the spot using the BT-Settl (Allard et al. 2012; Baraffe et al. 2015) stellar spectral model grids from the PHOENIX library (Husser et al. 2013). The spectral emission densities (SEDs) corresponding to the photosphere and the spot are governed by three fundamental stellar parameters: the stellar effective temperature ($T_{\rm eff}$), the stellar metallicity $[M/H]$, and the stellar surface gravity ($\log g$). For the purposes of this study, the metallicity and gravity are fixed at $[M/H] = 0$ (solar metallicity) and $\log g = 4.5$, respectively, in order to isolate the effects of active regions with contrasting temperatures. These are reasonable assumptions for low-resolution spectroscopy but may need to be reconsidered at higher resolution. In particular, the treatment of $\log g$ may need improvement, as it has been suggested that spots may be characterized by a lower $\log g$ than that of the photosphere due to the localized increase in magnetic pressure (e.g., Solanki 2003). In total, we require four SEDs, one corresponding to the photospheric temperature ($T_{\rm phot} = 4750$ K) and three corresponding to the spot temperatures considered in this work: 3750, 4250, and 4500 K, equivalent to a $\Delta T_{\rm spot}$ of 1000, 500, and 250 K, respectively.

Limb darkening is a well-known phenomenon acting to reduce the flux originating from the limbs of the stellar disk with respect to its center (Claret 2000; Howarth 2011). It also varies as a function of wavelength, and as such, different bands are characterized by different intensity profiles with the strongest effects seen in the optical. To account for the limb-darkening effect within our forward model we use the `ExoTETHyS` package, specifically the `SAIL` and `BOATS` subpackages (Morello et al. 2020a, 2020b, 2021), to calculate the limb-darkening coefficients (LDCs) for the star using the PHOENIX_2012_13 database (Claret et al. 2012, 2013) of BT-Settl models and the Claret four-coefficient law (Claret 2000):

$$\frac{I_\lambda(\mu)}{I_\lambda(1)} = 1 - \sum_{n=1}^{4} a_{n,\lambda}\,(1 - \mu^{n/2}), \tag{1}$$

where λ is the wavelength/bandpass being considered, $\mu = \cos\theta$ (where θ is the angle between the line of sight and the normal at the stellar surface), $I_\lambda(\mu)$ is the stellar intensity profile, $I_\lambda(1)$ is the intensity at the disk center (i.e., where $\mu = 1$), and $\alpha_{n,\lambda}$ are the LDCs.

We model the spot using the same LDCs that have been calculated for the star which is a reasonable approximation at first order, but again, may be revised for higher-resolution observations. The absolute fluxes of the star and the spot are also calculated using `ExoTETHyS` from which the normalized intensity profiles of the spotted star can be calculated as a function of wavelength (Figure 2). The emission from the spot-contaminated stellar disk is calculated as in Equation (2), where $S_{\rm star,\lambda}$, $S_{\rm phot,\lambda}$ and $S_{\rm spot,\lambda}$ are the spectra of the average star, the quiescent photosphere, and the spot, respectively, for a given wavelength (λ) and $F_{\rm spot}$ is the spot-filling factor. As such the resulting spectrum for the active star is essentially a combination of the photosphere and spot SEDs weighted by their relative covering fractions (Figure 3). The limb-darkening effect, which has already been defined for the 1000 annuli considered using Equation (1), is also accounted for in this stage.

$$S_{\rm star,\lambda} = ((1 - F_{\rm spot}) \times S_{\rm phot,\lambda}) + (F_{\rm spot} \times S_{\rm spot,\lambda}). \tag{2}$$

The resulting chromatic contamination can be described as acting as a contamination factor (ε) relative to the nominal transit depth i.e., the uncontaminated spectrum that would be observed in the case of an inactive star (Equation (3)). Where contamination is present only in the form of lower-temperature spots, as in this study, $\varepsilon_\lambda > 1$ resulting in increased transit depths at all wavelengths.

$$\varepsilon_\lambda = \frac{1}{1 - F_{\rm spot}\left(1 - \frac{S_{\lambda,\rm spot}}{S_{\lambda,\rm phot}}\right)}. \tag{3}$$

From the constructed, spot-contaminated stellar disks, we then use the open-source, light-curve analysis package `pylightcurve` (Tsiaras et al. 2016) to produce the spot-contaminated light curves (Figure 4) for all 200 wavelength intervals. These light curves are then used to construct the contaminated transmission spectrum for each spot case. We highlight that for this preliminary investigation, as we are only aiming to construct the contaminated transmission spectra for

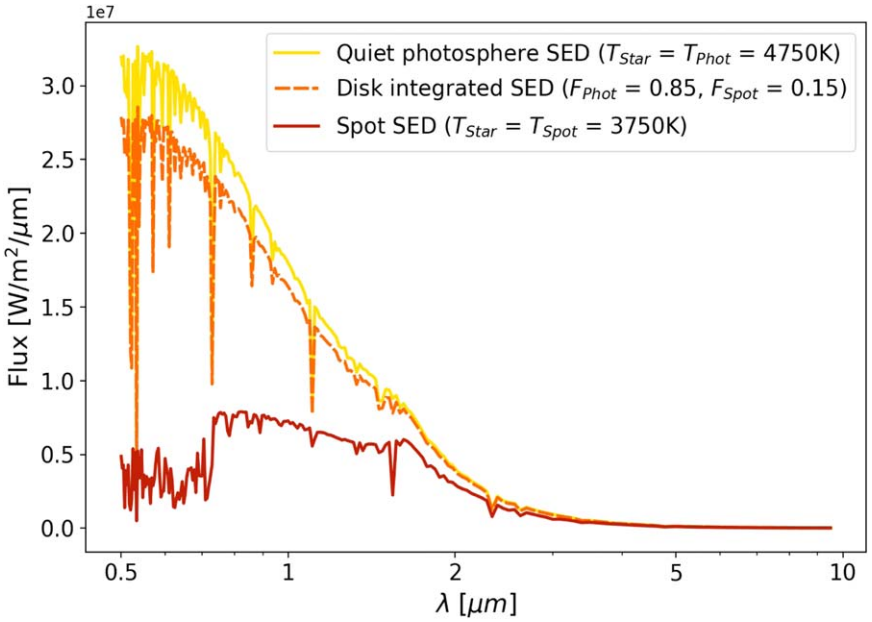

Figure 3. Three SEDs that are relevant to the construction of the forward stellar model. The two solid line SEDs correspond to the two temperature components that make up the surface of the active, heterogeneous stars considered in this work, the quiet photosphere, $T_{\rm phot} = 4750$ K (yellow), and the cooler starspot, $T_{\rm spot} = 3750$ K (red). The orange-dashed line SED represents the average, disk-integrated SED that would be observed in accordance with Equation (2) due to the weighted contributions of the quiet photosphere and the spot, assuming a spot-filling factor of $F_{\rm spot} = 0.15$ (and therefore a $F_{\rm phot} = 0.85$). We highlight that the spot-filling factor is varied throughout the spot-contaminated cases considered in this work (Table 1), as such, the disk-averaged SEDs will vary between cases.

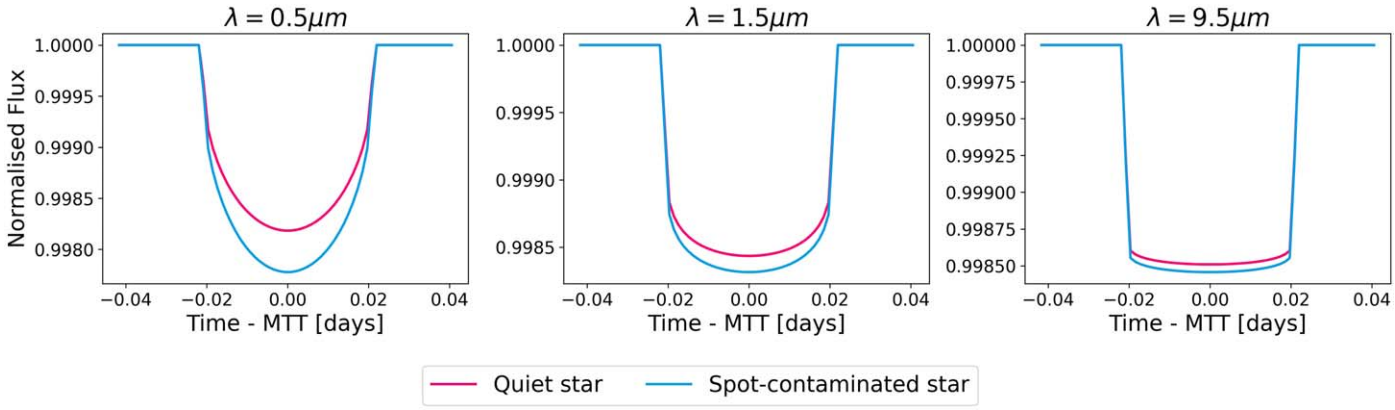

Figure 4. Uncontaminated (pink) and spot-contaminated (blue) light curves computed at 0.5 μm (left), 1.5 μm (center), and 9.5 μm (right) in the case of a large, central, high-contrast spot (Case 7 in Table 1).

use as inputs in the retrievals, we make the assumption that the orbital parameters used in `pylightcurve` are well constrained/known a priori and we do not introduce any Gaussian noise at the light-curve fitting stage to avoid introducing any additional bias that may arise arbitrarily. As in Section 2.1, error bars of 10 ppm are assumed for the input spectra. A comparison of the uncontaminated transmission spectrum and the spectrum with the highest degree of stellar contamination considered in this study is shown in Figure 5. The strongly wavelength-dependent nature of the spot contamination is evident, with it imparting a steep, positive blueward slope on the spectrum at wavelengths shortward of $\sim$2 μm. Longward of 2 μm, the contamination acts to introduce an almost constant positive offset equivalent to an overestimate in the transit depth on the order of 5% (>50 ppm) for the worst-case scenario.

2.4. The Retrieval Stellar Model: `ASteRA`

Retrievals on the grid of spot-contaminated spectra produced with the `StARPA` model are conducted using the fully Bayesian atmospheric retrieval code `TauREx3` (Al-Refaie et al. 2021). The main benefit of using `TauREx` for this study is that its modular design allows for a large amount of freedom and flexibility in testing and introducing new models. Mitigating for potential contamination due to stellar activity has been conducted in previous studies in the form of combined stellar-planetary retrievals (e.g., Pinhas et al. 2018; Bixel et al. 2019; Espinoza et al. 2019), and hopefully, this consideration of the host star will progressively become the new norm within the exoplanetary retrieval community. While this is not the first study focused on retrieval biases as a function of the degree of stellar heterogeneity, parameterized by temperature contrast and spatial extent (Iyer & Line 2020), this is the first time that that bias has been explored as a function of three spot parameters, with the inclusion of their position on the stellar disk, which governs how they interact with the limb-darkening effect.

The `ASteRA` plugin introduces a new, heterogeneous star class to `TauREx`, which follows a formalism similar to that

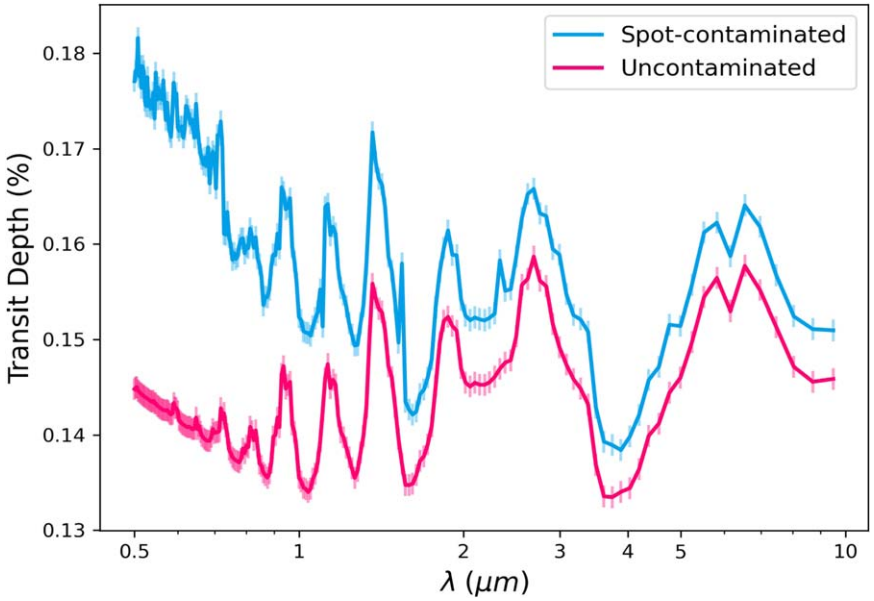

Figure 5. The effect of an extreme case of stellar contamination introduced by a large, high-contrast, unocculted, central spot (Case 7) on the observed transmission spectrum. The transit depths are overestimated at all wavelengths for the contaminated spectrum (blue) in comparison to the uncontaminated spectrum (pink). Error bars for both spectra are equivalent to 10 ppm. The magnitude of this overestimation increases exponentially as a function of decreasing wavelength with the strongest contamination seen in the optical regime.

used in previous stellar activity studies (e.g., Rackham et al. 2018, 2019). Instead of modeling the host star as being homogeneous and characterized by a single temperature/SED, the active star is modeled as a combination of multiple distinct temperature components, in this case, two: the quiet photosphere and the spot. The spot and photosphere SEDs are homogeneous disk-integrated models as they are not spatially resolved. The observed disk-integrated stellar spectrum is then produced by combining the two SEDs, weighted by the filling factor of each component, as shown in Figure 3 and described in Equation (2).

Using the heterogeneous star model requires the addition of two extra fitting parameters to a normal retrieval; the spot temperature $T_{\rm spot}$ and its filling factor $F_{\rm spot}$. For the purposes of this study, we restrict the active regions considered to spots only; however, `ASteRA` is easily extended to incorporate faculae, as has been done in the analysis of HST STIS observations of the hot Jupiter WASP-17b (Saba et al. 2022).

The core difference between the stellar models of `StARPA` and `ASteRA` is the treatment of limb darkening, or lack thereof. In contrast to the forward model produced with `StARPA`, the `ASteRA` stellar model is simpler in that the interplay between the spot and the limb-darkening effect is neglected, making it similar to the initial model used in Cracchiolo et al. (2021a). Conducting retrievals without the inclusion of limb darkening helps us gauge how important its inclusion is for low-resolution spectroscopy, particularly with respect to active host stars. With limb darkening neglected, the relative position of the spot on the stellar disk becomes unimportant provided that the entire spot is unocculted, reducing the dimensionality of the model by one. `ASteRA` requires the fundamental stellar parameters, i.e., effective temperature, metallicity, and $\log g$ as inputs in order to select the correct, corresponding PHOENIX spectra and these are fixed within the retrieval, as are the orbital parameters. This is done with the assumption that for real observations, these parameters will be reasonably accurately and homogeneously constrained a priori. Indeed, this is a high-priority and ongoing effort within the Ariel consortium regarding the stellar parameters (Danielski et al. 2022; Magrini et al. 2022) and the ExoClock project for the orbital parameters (Kokori et al. 2022).

Table 2
Fitting Parameters Used within the Retrievals with `ASteRA` and `TauREx` and Their Prior Bounds

Fitted Parameter	Prior Bounds	Scale
R_P ($R_{\rm Jup}$)	$0.75R_P$; $1.5R_P$	Linear
T_P (K)	100 ; 1000	Linear
$T_{\rm spot}$ (K)	3000 ; 4700	Linear
$F_{\rm spot}$	0 ; 0.99	Linear
$\log(H_2O)$	−12 ; −1	$\log_{10}$

For each spot case, two separate retrievals were conducted: one accounting for activity by fitting for $T_{\rm spot}$ and $F_{\rm spot}$ and one where the activity is not accounted for, despite being present. The planetary parameters R_P, T_P, and $\log(H_2O)$ are fit for in all retrievals and the prior bounds set for each parameter are given in Table 2. All retrievals conducted with `TauREx` use `MultiNest` (Feroz et al. 2009; Buchner et al. 2014) as the sampler with 500 live points and an evidence tolerance of 0.5. The retrievals with and without activity have a dimensionality of five and three, respectively. Conducting two retrievals allows for a comparison of the Bayesian evidence to obtain the Bayes factor, which quantifies how strongly the model with activity is favored over the one without. This provides an additional, statistically motivated way of quantifying how strong the spot contamination effects are. The retrieved values are then compared to the input/ground truth values for each parameter. From this, we determine how accurately the stellar and planetary parameters can be retrieved using `ASteRA` and if any bias is introduced as a result. The results of these retrievals are given in Section 3.

Table 3
Retrieved Planetary and Spot Parameters for Each of the 27 Spot Cases Investigated

Case No.	R_P ($R_{\rm Jup}$)	T (K)	log(H_2O)	Input $T_{\rm spot}$ (K)	Retrieved $T_{\rm spot}$ (K)	Input $F_{\rm spot}$	Retrieved $F_{\rm spot}$
GT	0.2730	400	−3.00	…	…	…	…
0	$0.2731^{+0.0003}_{-0.0002}$	$398.31^{+4.91}_{-5.62}$	-2.99 ± 0.05	N/A	$4670.15^{+20.12}_{-19.95}$	0	$0.391^{+0.400}_{-0.378}$
1	$0.2729^{+0.0004}_{-0.0005}$	$399.74^{+5.74}_{-6.36}$	-3.01 ± 0.05	3750	$3918.02^{+442.34}_{-512.02}$	0.010	$0.018^{+0.012}_{-0.004}$
2	$0.2732^{+0.0004}_{-0.0005}$	$395.98^{+5.58}_{-5.92}$	-2.99 ± 0.05	3750	$4017.22^{+490.05}_{-653.88}$	0.009	$0.015^{+0.030}_{-0.005}$
3	0.2736 ± 0.0003	$393.65^{+5.22}_{-6.00}$	-2.99 ± 0.05	3750	$4663.47^{+24.25}_{-585.54}$	0.005	$0.305^{+0.456}_{-0.299}$
4	0.2723 ± 0.0005	$403.97^{+5.68}_{-5.48}$	-3.05 ± 0.05	3750	$3779.57^{+112.28}_{-110.55}$	0.040	$0.059^{+0.004}_{-0.003}$
5	$0.2728^{+0.0004}_{-0.0005}$	$398.95^{+5.62}_{-5.94}$	-3.01 ± 0.05	3750	$3760.82^{+140.09}_{-139.26}$	0.035	$0.045^{+0.004}_{-0.003}$
6	0.2734 ± 0.0005	$392.42^{+6.08}_{-6.34}$	-2.96 ± 0.05	3750	$3773.34^{+449.70}_{-384.27}$	0.020	$0.018^{+0.007}_{-0.004}$
7	$0.2698^{+0.0004}_{-0.0005}$	$429.02^{+5.98}_{-5.45}$	-3.22 ± 0.05	3750	$3787.50^{+41.23}_{-59.60}$	0.160	$0.234^{+0.002}_{-0.006}$
8	$0.2711^{+0.0005}_{-0.0004}$	$411.12^{+5.23}_{-4.94}$	-3.09 ± 0.05	3750	$3707.40^{+39.81}_{-39.19}$	0.139	$0.176^{+0.003}_{-0.002}$
9	0.2737 ± 0.0004	$385.34^{+5.98}_{-6.09}$	-2.88 ± 0.05	3750	$3689.36^{+77.58}_{-85.94}$	0.080	0.070 ± 0.003
10	$0.2731^{+0.0004}_{-0.0005}$	$396.84^{+6.09}_{-5.72}$	-3.01 ± 0.05	4250	$4301.79^{+308.63}_{-744.13}$	0.010	$0.013^{+0.031}_{-0.006}$
11	$0.2733^{+0.0004}_{-0.0005}$	$396.01^{+5.58}_{-6.14}$	-3.01 ± 0.05	4250	$4534.88^{+139.69}_{-893.32}$	0.009	$0.018^{+0.491}_{-0.012}$
12	0.2733 ± 0.0003	$395.37^{+5.45}_{-5.64}$	-3.00 ± 0.05	4250	$4669.58^{+20.84}_{-48.39}$	0.005	$0.401^{+0.401}_{-0.387}$
13	0.2730 ± 0.0004	$398.71^{+5.72}_{-6.01}$	-3.02 ± 0.05	4250	$4517.02^{+97.03}_{-248.64}$	0.040	$0.100^{+0.092}_{-0.049}$
14	0.2731 ± 0.0004	$397.32^{+5.64}_{-5.79}$	-3.00 ± 0.05	4250	$4479.76^{+128.87}_{-277.08}$	0.035	$0.070^{+0.076}_{-0.035}$
15	0.2733 ± 0.0004	$394.98^{+5.48}_{-5.52}$	-2.98 ± 0.05	4250	$4299.29^{+288.83}_{-748.47}$	0.020	$0.015^{+0.031}_{-0.007}$
16	$0.2716^{+0.0006}_{-0.0005}$	$413.66^{+6.10}_{-7.66}$	$-3.13^{+0.06}_{-0.05}$	4250	$4311.13^{+72.60}_{-46.12}$	0.160	$0.221^{+0.044}_{-0.005}$
17	$0.2724^{+0.0007}_{-0.0005}$	$404.22^{+6.42}_{-8.11}$	-3.05 ± 0.06	4250	$4310.73^{+97.73}_{-63.42}$	0.139	$0.171^{+0.037}_{-0.020}$
18	$0.2736^{+0.0004}_{-0.0005}$	$391.16^{+6.20}_{-6.59}$	-2.95 ± 0.05	4250	$4341.04^{+235.08}_{-178.21}$	0.080	$0.073^{+0.146}_{-0.019}$
19	$0.2733^{+0.0003}_{-0.0004}$	$396.28^{+4.78}_{-5.62}$	$-3.01^{+0.04}_{-0.05}$	4500	$4657.20^{+29.61}_{-725.55}$	0.010	$0.145^{+0.592}_{-0.140}$
20	0.2733 ± 0.0003	$396.05^{+5.23}_{-5.85}$	-3.00 ± 0.05	4500	$4664.47^{+24.21}_{-452.27}$	0.009	$0.283^{+0.490}_{-0.278}$
21	0.2732 ± 0.0003	$396.10^{+5.37}_{-5.87}$	-2.99 ± 0.05	4500	$4670.68^{+19.06}_{-20.66}$	0.005	$0.404^{+0.402}_{-0.385}$
22	0.2729 ± 0.0004	$400.01^{+5.34}_{-6.01}$	-3.02 ± 0.05	4500	$4505.79^{+102.77}_{-276.72}$	0.040	$0.055^{+0.052}_{-0.028}$
23	0.2730 ± 0.0004	$398.58^{+5.35}_{-5.77}$	-3.01 ± 0.05	4500	$4456.58^{+149.41}_{-397.22}$	0.035	$0.033^{+0.047}_{-0.017}$
24	$0.2734^{+0.0003}_{-0.0005}$	$395.19^{+5.29}_{-5.78}$	-3.00 ± 0.05	4500	$4632.19^{+49.95}_{-842.62}$	0.020	$0.035^{+0.603}_{-0.030}$
25	0.2724 ± 0.0004	$406.09^{+5.52}_{-5.21}$	-3.07 ± 0.05	4500	$4570.62^{+57.93}_{-105.43}$	0.160	$0.400^{+0.02}_{-0.192}$
26	0.2728 ± 0.0004	$401.39^{+5.21}_{-5.64}$	-3.03 ± 0.05	4500	$4562.99^{+57.93}_{-105.43}$	0.139	$0.302^{+0.027}_{-0.149}$
27	$0.2733^{+0.0003}_{-0.0004}$	$398.53^{+4.93}_{-5.75}$	-2.98 ± 0.05	4500	$4501.57^{+111.42}_{-294.34}$	0.080	$0.065^{+0.066}_{-0.033}$

Note. The ground truth planetary parameters are given for comparison."Case 0" shows the parameters obtained when a retrieval using `ASteRA` is conducted on the uncontaminated spectrum as a frame of reference and to verify that `ASteRA` introduces no intrinsic bias.

3. Results

3.1. Retrievals Accounting for the Spot Contamination

In this section, we present the results of the retrievals for the 27 spot cases outlined in Section 2.2. To account for the spot contamination in retrieval, the spot parameters $T_{\rm spot}$ and $F_{\rm spot}$ are fitted for using the `ASteRA` plugin. In contrast, in Section 3.2, retrievals are conducted on the same contaminated spectra but with the incorrect assumption that the star is homogeneous. The same retrievals are also conducted on the uncontaminated transmission spectrum, which we term "Case 0". Case 0 acts as both a frame of reference for the highest precision and accuracy obtainable with `TauREx` for the simulated spectra used in this study, and as a verification that the use of `ASteRA` does not introduce any intrinsic bias. The results of these retrievals are given in Table 3 and a visual representation of the retrieval accuracy with respect to the planetary parameters is given in Figure 7. On inspection of the retrieved planetary parameters alone, the outlook is overall very positive, with the retrieved values falling very close to the ground truth in almost all of the cases considered. Intuitively, it becomes apparent that the largest errors in the planetary parameters are obtained for the cases where the spectra are most strongly contaminated. These cases are characterized by the largest, highest contrast spots considered in this study (e.g., Cases 7, 8, and 9). Despite showing the largest errors in this study, we emphasize that these errors are not substantial enough to result in a large misinterpretation of the planet. The retrieved parameters are also substantially more accurate than those obtained when the stellar contamination is neglected in the retrieval, the results of which are presented in the following section (Section 3.2). One other thing that becomes particularly evident here is that `ASteRA` struggles to constrain the spot-filling factor for small spots at high latitudes (Cases 3, 12, and 21) and for small, low-contrast spots (Cases 19, 20, and 21). For these scenarios, the retrieved $F_{\rm spot}$ is highly degenerate as the comparatively low levels of contamination can be reproduced by a larger number of spot configurations.

3.2. Retrievals When the Spot Contamination is Neglected

The retrieved parameters obtained for the 27 contaminated spectra when the spot parameters are not fit for are given in Table 4. In comparison to the retrievals presented in Section 3.1, the errors in the retrieved planetary parameters are far more significant, particularly for the highest-activity cases. For the most contaminated case (Case 7), the decision to account for stellar activity, even with the simplified method used by `ASteRA`, can be the difference in retrieving an approximately solar level water abundance ($\log(H_2O) = -3.22 \pm 0.05$) or an incorrect subsolar water abundance ($\log(H_2O) = -5.32 \pm 0.06$) an underestimation equivalent to

Table 4
Retrieved Planetary Parameters Obtained for the Same 27 Cases as in Table 3 but without Accounting for the Presence of Stellar Contamination by Fitting for the Spot Parameters

Case No.	R_P ($R_{\rm Jup}$)	T (K)	log(H_2O)
GT	0.2730	400	−3.00
0	0.2731 ± 0.0002	$398.40^{+5.08}_{-5.50}$	-3.00 ± 0.05
1	$0.2746^{+0.0003}_{-0.0002}$	$390.68^{+5.63}_{-6.30}$	-3.09 ± 0.05
2	$0.2743^{+0.0003}_{-0.0002}$	$389.63^{+6.06}_{-5.93}$	$-3.05^{+0.05}_{-0.04}$
3	$0.2736^{+0.0003}_{-0.0002}$	$392.96^{+5.61}_{-5.76}$	$-3.00^{+0.04}_{-0.05}$
4	0.2792 ± 0.0003	$365.30^{+6.97}_{-7.10}$	-3.41 ± 0.05
5	0.2781 ± 0.0003	$368.64^{+6.72}_{-6.93}$	-3.27 ± 0.05
6	0.2753 ± 0.0003	$381.90^{+5.82}_{-6.23}$	-3.05 ± 0.05
7	0.2986 ± 0.0001	$274.94^{+0.18}_{-0.28}$	-5.32 ± 0.06
8	0.2944 ± 0.0001	$274.93^{+0.29}_{-0.41}$	-4.44 ± 0.05
9	0.2827 ± 0.0003	$330.52^{+7.90}_{-7.58}$	-3.28 ± 0.07
10	$0.2738^{+0.0003}_{-0.0002}$	$393.36^{+5.17}_{-5.95}$	-3.04 ± 0.05
11	0.2737 ± 0.0002	$393.43^{+5.35}_{-5.80}$	-3.03 ± 0.05
12	$0.2733^{+0.0003}_{-0.0002}$	$395.31^{+5.07}_{-6.02}$	-3.00 ± 0.05
13	0.2758 ± 0.0003	$385.69^{+6.47}_{-6.23}$	-3.24 ± 0.05
14	0.2753 ± 0.0003	$386.20^{+6.24}_{-6.26}$	-3.16 ± 0.05
15	0.2741 ± 0.0001	$390.73^{+5.56}_{-6.18}$	$-3.04^{+0.05}_{-0.04}$
16	$0.2833^{+0.0004}_{-0.0003}$	$374.04^{+7.45}_{-7.97}$	$-4.16^{+0.05}_{-0.04}$
17	0.2819 ± 0.0003	$363.24^{+7.63}_{-7.01}$	-3.80 ± 0.05
18	0.2774 ± 0.0003	$370.91^{+6.50}_{-6.51}$	-3.21 ± 0.05
19	0.2735 ± 0.0002	$395.05^{+5.17}_{-5.87}$	$-3.02^{+0.05}_{-0.04}$
20	$0.2734^{+0.0003}_{-0.0002}$	$395.45^{+5.07}_{-5.87}$	-3.01 ± 0.05
21	0.2732 ± 0.0002	$396.01^{+5.06}_{-5.61}$	$-2.99^{+0.04}_{-0.05}$
22	$0.2745^{+0.0003}_{-0.0002}$	$395.50^{+5.09}_{-6.24}$	-3.13 ± 0.04
23	0.2742 ± 0.0002	$392.38^{+5.42}_{-5.78}$	-3.09 ± 0.04
24	$0.2736^{+0.0003}_{-0.0002}$	$393.47^{+5.08}_{-5.78}$	$-3.01^{+0.04}_{-0.05}$
25	$0.2784^{+0.0003}_{-0.0002}$	$384.85^{+5.39}_{-6.03}$	$-3.61^{+0.05}_{-0.04}$
26	0.2775 ± 0.0003	$382.73^{+7.00}_{-6.67}$	-3.44 ± 0.05
27	0.2752 ± 0.0003	$385.92^{+6.21}_{-6.34}$	-3.13 ± 0.05

Note. Comparison with the ground truth shows that in cases of severe activity (e.g., Cases 7, 8, and 9) the planetary parameters retrieved are highly inaccurate.

>2 orders of magnitude. The planetary temperature (T_P) is also significantly underestimated with the retrieved temperature of 275 K being 125 K cooler than the ground truth (400 K), which, in the context of a temperate planet, represents a very substantial error (>30%).

Conducting two retrievals on the same contaminated spectrum allows us to determine which model is preferred by the data through the comparison of the Bayesian evidence. The Bayes factors (*lnB*) show a strong preference for the ASteRA retrieval for the majority of cases considered here (Figure 6). The cases in which only a weak preference for the corrected model is seen, or a preference for the retrieval in which contamination is neglected is indicated, correspond to those characterized by small, high-latitude, and low-temperature contrast spots. As the contamination resulting from such spots is minimal, the model neglecting contamination is still able to perform reasonably well under conditions of low activity. In contrast, a penalty is applied to the model accounting for contamination when calculating the Bayesian evidence as a result of its higher dimensionality, explaining why the activity model is not conclusively preferred in the low activity regime despite a spot being present. For these low-activity cases, the planetary parameters are still accurately retrieved even when the presence of the spot is neglected entirely (Figure 7). The Bayes factor shows the strongest preference for the retrieval without the activity correction for the uncontaminated case as one would expect. This reaffirms that no intrinsic bias is introduced by ASteRA and that ASteRA does not find evidence for stellar activity where there is none.

4. Discussion

4.1. Understanding the Interaction between Spot Contamination and the Limb-darkening Effect

In order to adequately understand and interpret the retrieval results we must first consider how each of the three spot parameters will contribute to the contamination factor and therefore the observed spectrum. The size and the temperature contrast of the spot have very intuitive impacts on the contamination spectrum. The contamination factor increases as a function of both as expected. In comparison, the effect of the spot's latitude/position on the stellar disk is more subtle. It also affects only the shortest wavelengths with the largest variations seen at $\lambda \leqslant 1\mu$m. The importance of the spot latitude in this study is twofold. First, the projected filling factor (defined as the percentage of the 2D stellar disk that is covered by the spot) decreases as a function of latitude for a spot of a constant radius due to decreases in its 3D geometric projection onto the surface. This reduces the contamination introduced by the spatial coverage of the spot alone. Second, the position of the spot dictates how it will interact with the limb-darkening effect as shown by the normalized intensity profiles (Figure 2) in Section 2.3. We subsequently term this interaction as the limb darkening–spot interplay. The limb darkening–spot interplay is an important consideration as a central spot will remove a greater amount of flux compared to a spot (with an identical filling factor) located at the limb as the intensity maximum occurs at the disk center.

4.2. Accuracy of the Retrieved Planetary Parameters

In this section, we focus on how the use of the simplified spot model within ASteRA affects the retrieved planetary parameters (Figure 7) and how these errors are observed to vary as a function of the spot parameters. As already stated in Section 3.1, the largest errors are seen for the most extreme activity cases. Figure 7 allows for the visual comparison of the values retrieved when the spot is fit for (black data points) as opposed to when it is not (red data points). This reiterates the large improvement in accuracy obtained through using even a simple activity correction method over no correction at all. In the worst-case scenario not correcting for stellar activity leads to an underestimation of the water mixing ratio by over two orders of magnitude. Such a large error would significantly impact our understanding of the planet as a whole. Another concerning result highlighted by Figure 7 is that, while the bias introduced by not correcting for stellar activity strongly affects the retrieval accuracy, it does not affect the retrieval precision, as evidenced by the small error bars. As such this high precision could be very misleading for real observations where a priori knowledge of the correct planetary parameters is not known. This is consistent with previous retrieval-oriented works, e.g., Iyer & Line (2020) and reaffirms that caution should be taken when analyzing retrieval results where stellar contamination has not been included, even if the host star is thought to be less active.

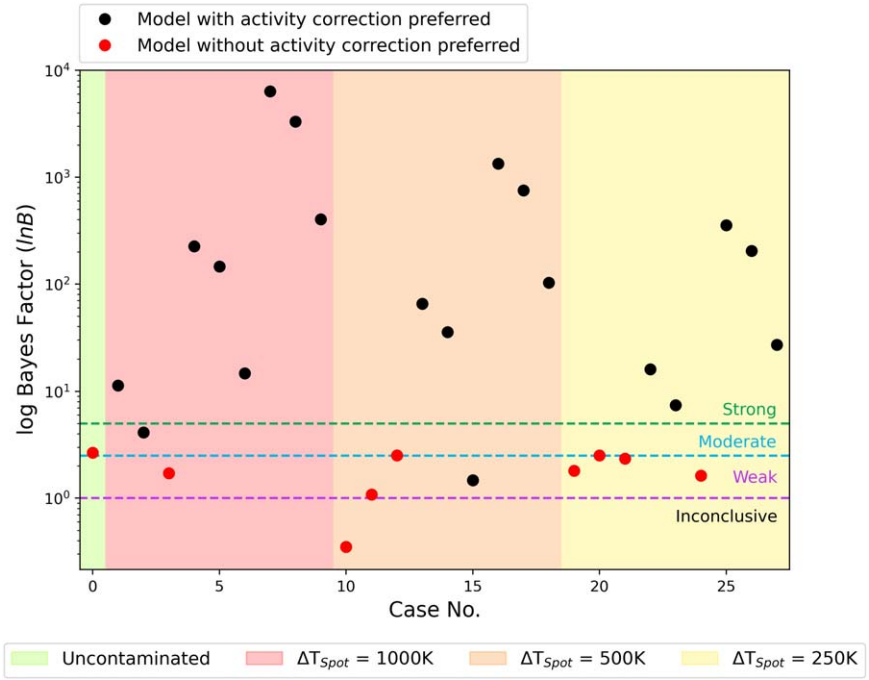

Figure 6. A graphical representation of the Bayes factors for the uncontaminated and 27 spot-contaminated cases explored. Black markers indicate that the Bayes factor is in favor of the ASteRA retrieval, where stellar activity has been accounted for and corrected for. Red markers indicate a preference for the lower dimensionality retrieval in which the spot parameters are not fit for. The Jeffreys scale (Trotta 2008) is overplotted to show the strength of the model preference where a Bayes factor $\geqslant$1 is indicative of weak preference, $\geqslant$2.5 of moderate preference, and $\geqslant$5 a strong preference. A Bayes factor below 1 is deemed to be inconclusive.

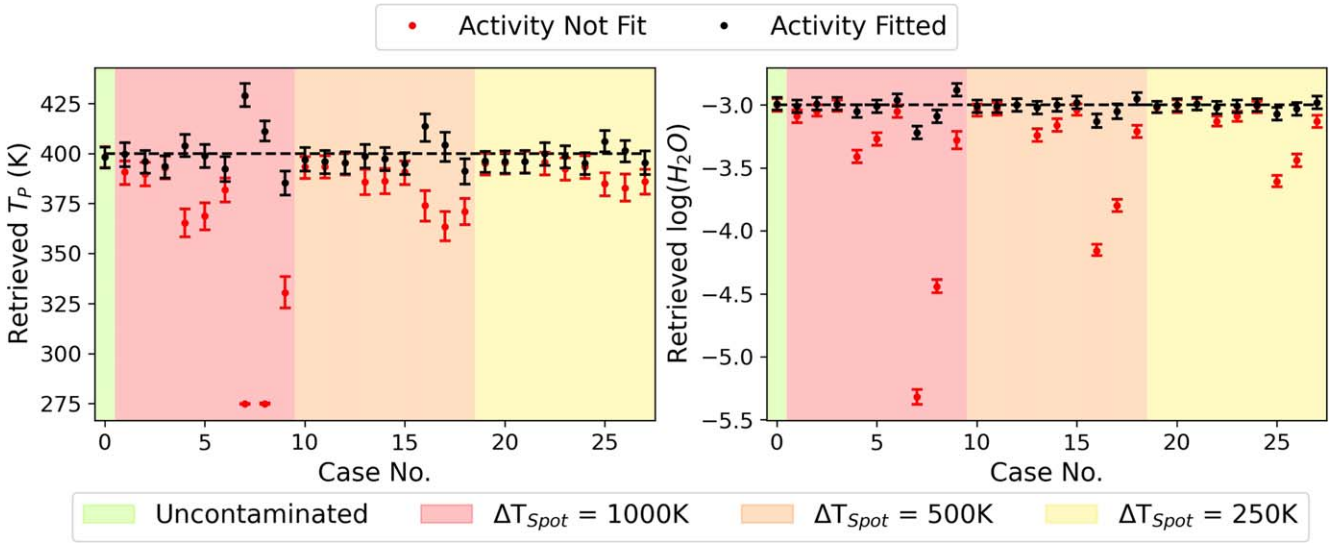

Figure 7. The retrieved planetary temperature (T_P; left) and H_2O mixing ratio (log(H_2O); right) obtained for each spot case when the spot parameters are fit for (black data points) vs. when activity is not accounted for (red data points). The plots' background colors correspond to the temperature contrasts of the spot considered for each case. The ground truth for each parameter is indicated by the black-dashed line.

The parameter that appears to be most influential in the highest-activity cases is the spot latitude, which acts as a proxy for limb darkening (Figure 8). It is intuitive that this spot parameter should have the greatest effect on the residual bias as ASteRA has no way of fitting for the spot's position. A decreasing trend as a function of increasing $\phi_{\rm spot}$ is observed in the retrieved planetary temperature, with this being overestimated for the lower-latitude spot cases (Cases 7 and 8) and subsequently underestimated at the highest latitude considered (Case 9). This underestimation can be attributed to the interplay of a high-latitude spot and limb darkening. The reduced flux originating from the quiet photosphere at the limbs acts to reduce the observed spot contrast and thus there is also a reduction in the degree of contamination introduced. In contrast to this, the opposite trend is observed in the retrieved H_2O mixing ratio with this being underestimated at the two lower latitudes and overestimated at the highest. Similar trends in T and log(H_2O) are observed for the other cases considering large (0.4 R_*) spots (Cases 16, 17, and 18 and Cases 25, 26, and 27 respectively); however, the magnitude of the bias introduced is weaker due to the lower-spot contrasts considered.

4.3. Accuracy of the Retrieved Stellar Parameters

ASteRA is less successful in recovering the spot parameters (Figure 9). This is likely due to a combination of not

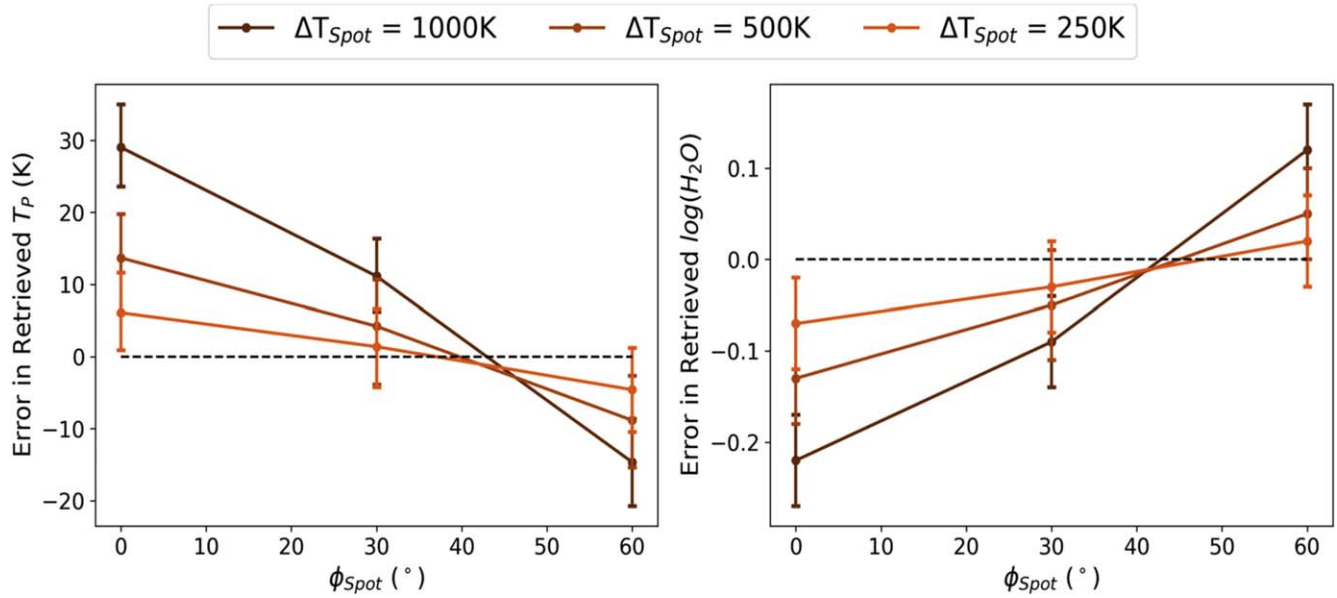

Figure 8. The error introduced in the retrieved planetary temperature T_P (left) and atmospheric H_2O mixing ratio log(H_2O) (right) as a function of spot latitude. Only the cases with the largest spots (0.4 R_*) are plotted as these are the cases where some minor residual contamination remains after the correction. The effect of the spot temperature contrast on the observed trend is evident with the strongest trend seen for the highest contrast spots. There is a clear anticorrelation between T_P and log(H_2O).

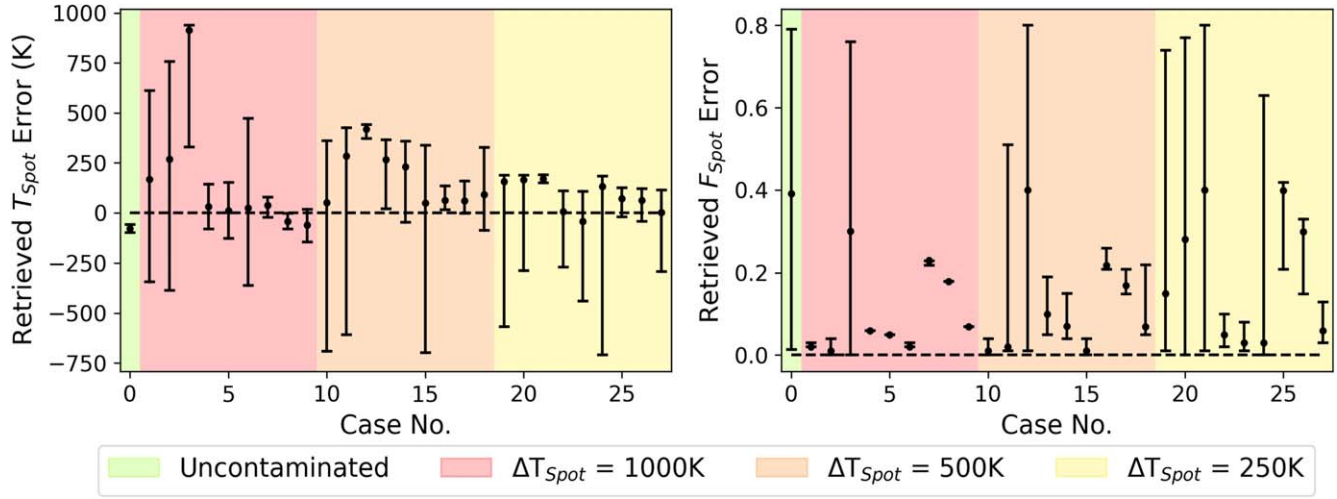

Figure 9. The error in the retrieved spot parameters $T_{\rm spot}$ and $F_{\rm spot}$ obtained for the uncontaminated and the 27 spot-contaminated cases. Figure elements are the same as in Figure 7. Note that $F_{\rm spot}$ is very poorly constrained for the uncontaminated case, which is consistent with the absence of stellar activity. In contrast, $T_{\rm spot}$.

accounting for limb darkening and also because many combinations of the three spot parameters are degenerate at low resolution, particularly if the spot in question has a low-to-moderate temperature contrast. An example of how different spot parameter combinations can result in degenerate solutions is given in Figure 10 for clarity. The largest errors and uncertainties in the retrieved $T_{\rm spot}$ are seen for the smallest spots considered, especially when present at high latitudes. For larger spots, $T_{\rm spot}$ is generally reasonably well constrained due to the larger contamination effects they introduce. Large errors and uncertainties are also seen in the retrieved $F_{\rm spot}$ values in several cases. Large error bars point to substantial degeneracy in the small spot cases, whereas in the case of the largest spots, $F_{\rm spot}$ is often constrained to a higher precision but significantly overestimated. The results of these retrievals indicate that we should be more cautious with retrieved stellar parameters, as the degeneracies between them mean that they are constrained less confidently by the retrieval. Simultaneous, external observations, e.g., at different bandpasses could help break some of these degeneracies.

4.4. The Effect of Limb Darkening

In order to attribute the biases seen in the planetary parameters to not accounting for the effect of limb darkening, the worst-case scenarios (Cases 7, 8, and 9) were rerun. For each case, the contaminated spectrum was regenerated with the LDCs set to zero (Figure 11) and a further retrieval was conducted. When noise is taken into account the effect of including versus excluding the limb-darkening effect is really only distinguishable at the shortest wavelengths considered ($\lambda < 1\,\mu$m), even for these worst-case scenarios. Ariel in particular will only be able to access this wavelength region through three photometric bands (Tinetti et al. 2021), as such,

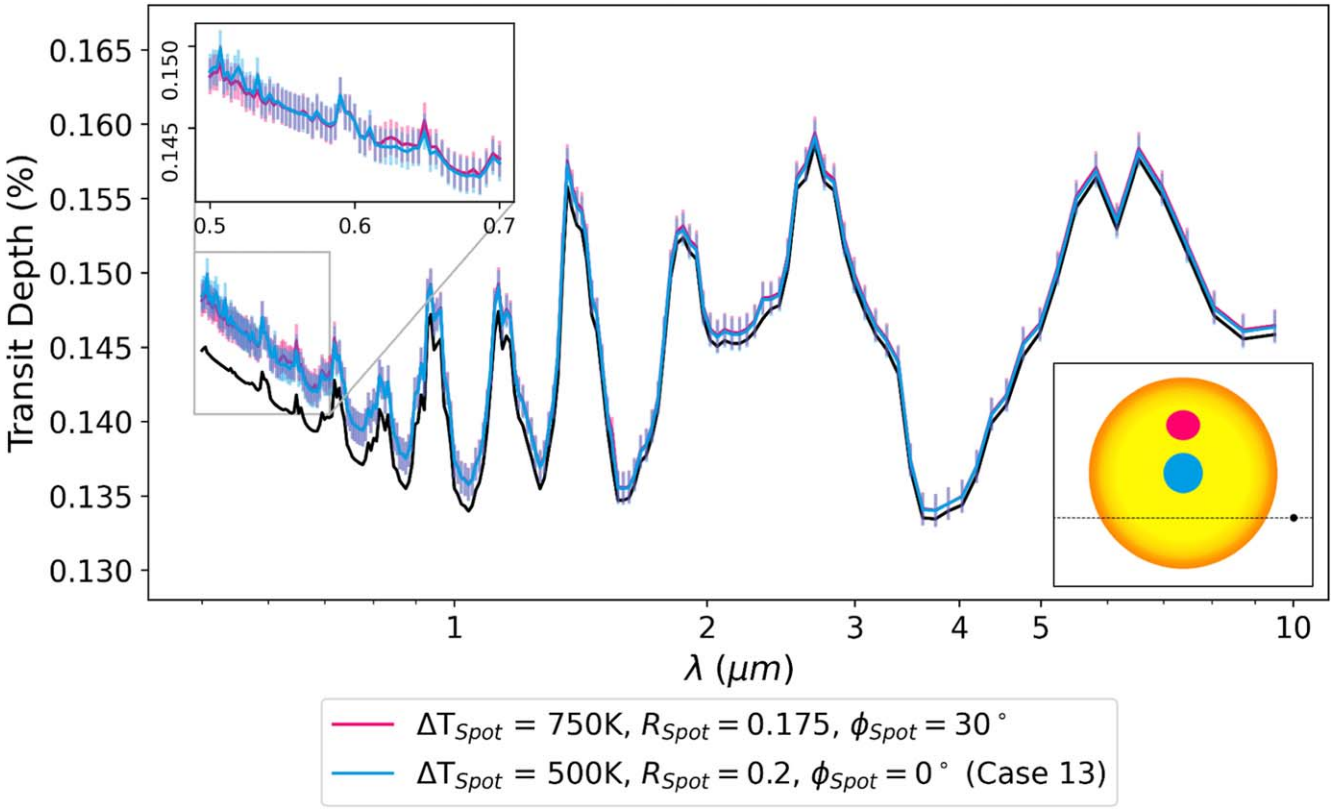

Figure 10. An example of how spot parameters can be degenerate at low resolution, particularly in cases of moderate activity. The black line depicts the uncontaminated transmission spectrum. In blue is the contaminated spectrum obtained with a slightly larger ($R_{\rm spot} = 0.2R_*$), central ($\phi_{\rm spot} = 0°$) spot with a lower temperature contrast ($\Delta T_{\rm spot} = 500$ K), which is equivalent to Case 13 in our retrieval grid. In contrast to this the pink contaminated spectrum results from a smaller ($R_{\rm spot} = 0.175R_*$), higher-latitude ($\phi_{\rm spot} = 30°$), and higher-contrast ($\Delta T_{\rm spot} = 750$ K) spot. The two-spot-contaminated spectra are undifferentiable beneath the noise at all wavelengths. As such, a retrieval would not be able to confidently differentiate between these two solutions. Inset (top left) A close-up view of the two-spot-contaminated spectra at 0.5–0.7 μm where they diverge the most. The differences between them are still easily lost beneath the very optimistic noise estimate considered here (10 ppm). Inset (bottom right) A visual depiction of the two degenerate spots.

extracting as much information as possible about the activity of the host star from these three data points will be of utmost importance. The retrieved parameters when the limb-darkening effect is removed are given in Table 5. With the exception of $T_{\rm spot}$, which was already well constrained, all other parameters are retrieved more accurately, including $F_{\rm spot}$ now being constrained correctly. As the input observations are inherently different due to the removal of the limb-darkening effect, unfortunately, we cannot use the Bayes factor as a means of model comparison here as was done in previous sections (Figure 6). However, the fact that both the planetary and stellar parameters are retrieved more accurately when the limb-darkening contribution is removed from the input observation provides compelling evidence in itself that the residual bias seen originates from the limb darkening–spot interplay. For real observations ignoring the effects of limb darkening within activity correction frameworks will unfortunately not be a viable solution as this phenomenon will always be present. As such, we believe that as a community there is a need for us to push toward developing and using stellar activity models in which the limb darkening–spot interplay is accounted for, particularly when dealing with highly active host stars.

The posterior distributions retrieved for the two input scenarios, spot contaminated with the limb-darkening contribution and spot contaminated with the limb-darkening contribution removed, are given in Appendix A. The greatest deviation between the two posteriors is seen in the case of a central spot. This is intuitive as the spot masks the disk center where a greater proportion of the stellar flux originates from and as such, results in the largest residual contamination due to not encompassing the limb darkening–spot interplay. As the spot is modeled at progressively higher latitudes, less residual contamination remains and the posteriors for the limb darkened, contaminated spectra converge toward the correct values.

4.5. Preliminary Investigations into Multiple-spot Cases

In the previous sections, we have focused only on isolated, single spots in order to assess the interactions of the spot parameters and how they influence the observed, contaminated spectrum. In order to better understand the limb darkening–spot interplay (Section 4.1) and the bias it introduces in retrievals (Section 4.4) we conducted three preliminary multiple-spot cases using the same methodology to explore how this interplay may differ when more than one spot is present. For comparability, the spots have the same temperature ($T_{\rm spot} = 3750$ K) and total filling factor ($F_{\rm spot} = 0.16$) as the single-spot Case 7. The multiple-spot cases we consider are a two-spot case and two cases characterized by 10 smaller spots. These 10 spots are arranged in a random configuration on the stellar disk in the first scenario, and occupy a preferred active latitude centered around $\phi = 15°N$ in the second scenario. The only strict requirement for these cases is that all of the spots remain unocculted. The resultant contaminated spectra are shown in

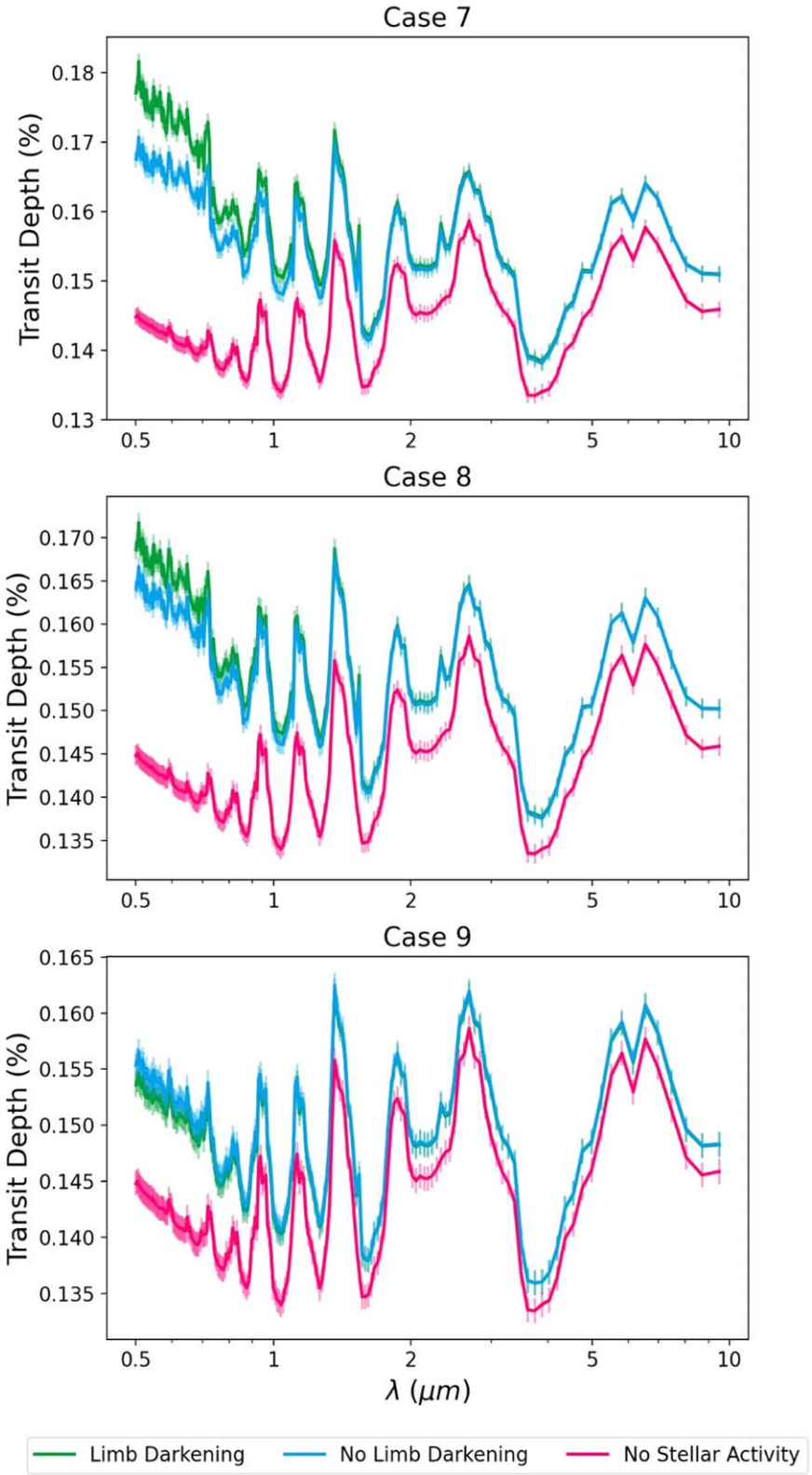

Figure 11. Forward model transmission spectra highlighting the effect of accounting for the contribution of limb darkening when considering the spot contamination or not for the most severely contaminated cases (Cases 7, 8, and 9). The uncontaminated spectrum assuming a quiet host star is shown in pink. The green spectrum is the spot-contaminated spectrum produced when the contribution of the limb-darkening effect to the overall contamination is considered, the blue spectrum, in contrast, is the contaminated spectrum obtained when limb darkening is not accounted for. The error bars are equivalent to 10 ppm.

Figure 12 alongside the uncontaminated spectrum and the Case 7 contamination spectrum, with and without the contribution of the limb-darkening effect, as in Figure 11. These spectra show that, as the number of spots increases and they are distributed over a larger range of μ values (i.e., located at different distances from the disk center) the limb darkening–spot interplay and its contribution to the contamination spectrum are minimized. A table of the retrieved parameters is given in

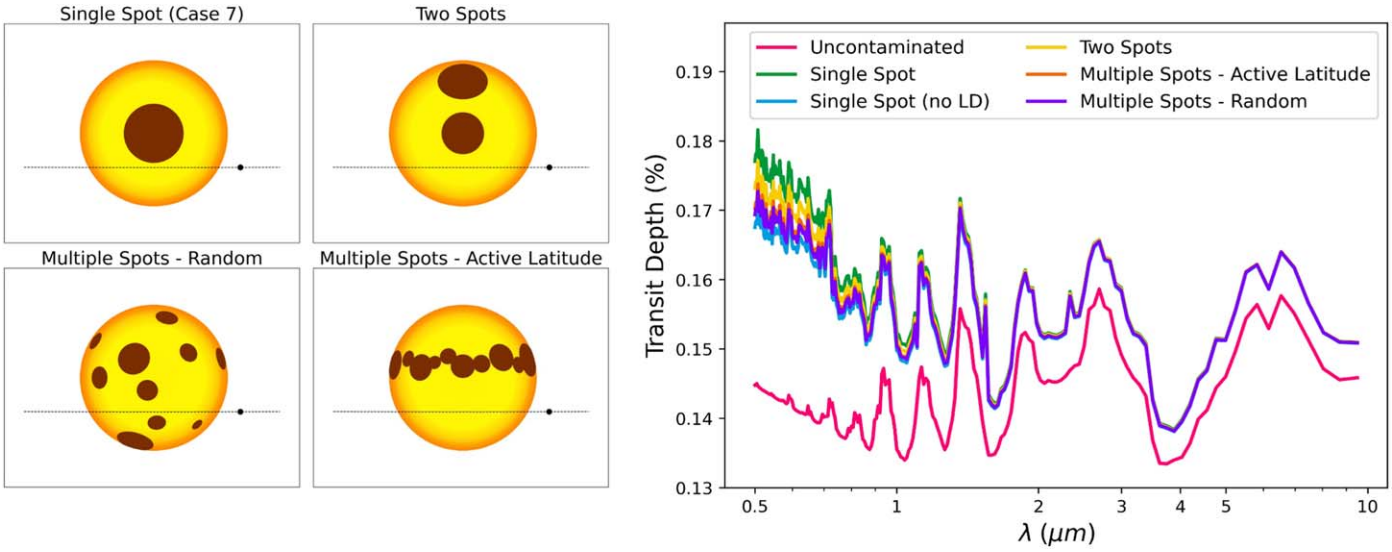

Figure 12. Left: visual representations of the single-spot Case 7 and the three multiple-spot cases investigated: a two-spot case, a multiple-spot case with a random spot configuration, and a multiple-spot case where the spots occupy a preferred active latitude centered around $\phi = 15°$. All spots have the same temperature contrast ($\Delta T_{\rm spot} = 1000$ K) and their combined filling factors are kept constant at $F_{\rm spot} = 0.16$. Right: the resulting contaminated spectra for the two-spot (yellow), active latitude (orange), and random configuration (purple) multiple-spot cases compared to the uncontaminated spectrum (pink) and the single-spot of Case 7 with (green) and without (blue) the contribution of the limb-darkening effect. The 10 ppm error bars have been removed here for clarity.

Table 5
The Retrieved Spot and Planetary Parameters Obtained for Cases 7, 8, and 9 When the Effect of Limb Darkening is No Longer Present in the Input Spectra i.e., the Blue Spectra in Figure 11

Case No.	$T_{\rm spot}$	Input $F_{\rm spot}$	Retrieved $F_{\rm spot}$	R_P ($R_{\rm Jup}$)	T	$\log(H_2O)$
GT	3750	⋯	⋯	0.273	400	−3
7	$3700.81^{+34.68}_{-33.72}$	0.160	0.166 ± 0.002	0.2728 ± 0.0004	$392.47^{+5.69}_{-6.06}$	-2.91 ± 0.05
8	$3704.40^{+33.81}_{-38.06}$	0.139	0.144 ± 0.002	0.2728 ± 0.0004	$393.19^{+6.37}_{-5.51}$	-2.92 ± 0.05
9	$3715.14^{+82.23}_{-58.94}$	0.080	0.084 ± 0.003	0.2730 ± 0.0004	$395.72^{+5.58}_{-6.15}$	-2.95 ± 0.05

Note. The ground truth spot and planetary parameters are given for comparison. $F_{\rm spot}$ varies on a case-by-case basis, as such, the Input $F_{\rm spot}$ denotes the ground truth value for each case.

Table 6
The Retrieved Planetary and Spot Parameters for the Multiple-spot Cases Considered in Section 4.5 Compared to the Ground Truth Values and Case 7, a Single Spot of the Same Temperature and Filling Factor

	R ($R_{\rm Jup}$)	T (K)	$\log(H_2O)$	$T_{\rm spot}$ (K)	$F_{\rm spot}$
Ground truth	0.273	400	−3	3750	0.16
Case 7	$0.2698^{+0.0004}_{-0.0005}$	$429.02^{+5.98}_{-5.45}$	-3.22 ± 0.05	$3787.50^{+41.23}_{-59.60}$	$0.234^{+0.002}_{-0.006}$
Two spot	0.2705 ± 0.0004	416.80 ± 5.10	$-3.11^{+0.04}_{-0.05}$	$3702.22^{+37.76}_{-36.41}$	0.203 ± 0.002
Multiple spot—random	0.2721 ± 0.0004	$400.92^{+5.81}_{-6.68}$	-2.97 ± 0.05	$3701.30^{+34.30}_{-34.57}$	0.178 ± 0.002
Multiple spot—active latitude	0.2716 ± 0.0004	$405.52^{+5.07}_{-5.58}$	-3.01 ± 0.05	$3700.95^{+36.00}_{-34.57}$	0.185 ± 0.002

Appendix B, alongside the posterior distributions for the three multiple-spot cases considered.

These additional multiple-spot experiments highlight that the large, high-contrast single spots considered as the main focus of this study likely do represent the worst-case scenario when considering the additional complication arising from the limb-darkening effect. Observationally, the regimes in which the limb darkening–spot interplay will really start to matter will therefore be those in which the system geometries most closely replicate those in Cases 7 and 8. Large high-latitude and polar spots have been proposed to exist on several stars (e.g., Järvinen et al. 2018; Almenara et al. 2022; Strassmeier et al. 2023) although any biases that these could introduce will be lessened if the star's rotation axis is not significantly inclined with respect to the line of sight as the spot will appear at the limb. The worst-case scenario would occur for a system where the host star both possesses a polar spot and is inclined relative to the transiting planet and observer. In such a scenario the polar spot could manifest as a large central spot. Examples such as the HAT-P-11 system, which consists of an active K4-dwarf hosting two highly misaligned planets (e.g., Sanchis-Ojeda & Winn 2011; Morris et al. 2017; Yee et al. 2018) and the almost pole-on, solar-type star τ Ceti, which is frequently included in target lists for exoplanet searches (Korolik et al. 2023) both indicate that observing a system with such geometry in the future cannot be ruled out.

Table 7
The Input Umbra and Penumbra Parameters for the Two Radial Temperature Variation Spot Cases Considered in Section 4.6

	$T_{\rm Umbra}$ (K)	$R_{\rm Umbra}$ (R_*)	$F_{\rm Umbra}$	$T_{\rm Pen}$ (K)	$R_{\rm Pen}$ (R_*)	$F_{\rm Pen}$
Temperature variation Case 1	3750	0.3	0.09	4250	0.4	0.07
Temperature variation Case 2	3750	0.2	0.04	4250	0.4	0.12

Table 8
Retrieved Planetary and Spot Parameters for the Two Radial Temperature Variation Spot Cases Considered in Section 4.6 Compared to the Highest-activity, Single-temperature Spot Case (Case 7) and the Ground Truth Values

	R ($R_{\rm Jup}$)	T (K)	$\log(H_2O)$	$T_{\rm spot}$ (K)	$F_{\rm spot}$
Ground truth	0.273	400	−3	…	…
Case 7	$0.2698^{+0.0004}_{-0.0005}$	$429.02^{+5.98}_{-5.45}$	-3.22 ± 0.05	$3787.50^{+41.23}_{-59.60}$	$0.234^{+0.002}_{-0.006}$
Temperature variation Case 1	$0.2721^{+0.0004}_{-0.0003}$	$446.52^{+5.57}_{-5.66}$	-3.39 ± 0.04	$4201.09^{+33.84}_{-34.46}$	0.254 ± 0.003
Temperature variation Case 2	$0.2715^{+0.0003}_{-0.0004}$	$440.65^{+5.86}_{-5.84}$	$-3.33^{+0.05}_{-0.04}$	$4299.21^{+33.98}_{-33.68}$	0.278 ± 0.004

Our recommendation for dealing with this worst-case, high-activity regime, where it will be necessary to include the limb darkening–spot interplay in order to fully negate the bias introduced by stellar contamination, is to improve our retrieval stellar models so that they are capable of also parameterizing the position of spots on the stellar disk. This is the focus of ongoing work, however, as with increasing the dimensionality of any model, we cannot ignore the risk of injecting intrinsic bias if the additional complexity is not physically motivated. For real observations, there will also be the added challenge of not having any a priori knowledge of spot positions as the stellar disks are not spatially resolved. This work has shown that at first order using the simpler, two-parameter spot prescription as is done with `ASteRA` is sufficient to remove the majority of the bias and enable a good understanding of the planet's atmospheric properties. A push toward a better understanding of the limb-darkening effect, whether in the presence of stellar activity or not, will also be highly beneficial. This will be especially important for later-type stars where offsets due to the choice of limb-darkening treatment appear to be inevitable (Patel & Espinoza 2022).

4.6. Preliminary Investigation into Spots Displaying a Radial Temperature Variation—A Separation into Umbra and Penumbra

All of the retrievals presented in the previous sections have been conducted assuming a single $T_{\rm spot}$. In order to explore the validity of this assumption, we conducted preliminary studies into modeling the contamination introduced by spots with radial temperature variations. This section examines the resulting impact on retrieval performance. We assume the same spot geometry as was used in the highest contamination, worst-case scenario (Case 7). The spot is now separated into a cooler central umbra characterized by a temperature $T_{\rm umbra} = 3750$ K, and a surrounding penumbra characterized by a temperature $T_{\rm pen} = 4250$ K. This separation into two regions of distinct temperatures is consistent with observed sunspots and with the 3D MHD models for later-type stars produced by Panja et al. (2020). We explore two different cases in which the total spot-filling factor ($F_{\rm spot}$) is kept constant but the relative fractions of the umbra ($F_{\rm umbra}$) and penumbra ($F_{\rm pen}$) are varied. Table 7 shows the input parameters for these two cases. In doing this, we explore the retrieval response to contamination effects resulting from a spot that is characterized by two different temperatures/SEDs when the retrieval model is only capable of fitting this contamination using a single $T_{\rm spot}$ value.

As expected, this further discrepancy between the forward model and the retrieval model introduces an additional source of error to the retrieved planetary parameters. The retrieved parameters are given in Table 8 alongside the equivalent single $T_{\rm spot}$ case (Case 7) for comparison. The posterior distributions for the two-spot temperature variation retrievals are given in Appendix C. The retrieved spot parameters provide insight into how the retrieval has attempted to correct for the dual temperature spot. In both cases, the retrieval is best able to fit for the contamination with a moderate $T_{\rm spot}$ close to $T_{\rm pen}$, and an overestimated $F_{\rm spot}$. The overestimation of the spot-filling factor accommodates the increased contamination due to the higher-contrast umbra. The posteriors show an increasing correlation and degeneracy between $T_{\rm spot}$ and $F_{\rm spot}$ compared to the single-$T_{\rm spot}$ cases. The introduction of radial temperature variation slightly enhances the biases in the retrieved planetary parameters that were observed for the highest-activity cases. Intuitively, the greater effect is seen when considering a spot with a higher umbra–penumbra ratio, resulting in an overestimation of the retrieved planetary temperature of ∼47 K and an underestimation of the water mixing ratio of 0.39 magnitude. Although this error is non-negligible, importantly it is not limiting. Retrievals conducted in the presence of such contamination would still be useful and informative.

5. Conclusion

The main objective of this paper is to determine how complex our stellar activity models should be in order to remove biases so that we can characterize the transiting planet as accurately and efficiently as possible. At the same time, we want to ensure that our retrieval analysis remains reliable. As such, it is important that we do not introduce an unjustified complexity that may inject a bias intrinsically. We make use of a grid of 27 spot-contaminated stellar disks created with the more complex forward stellar model (`StARPA`), in which the interplay of the spot and limb darkening is accounted for, and conduct retrievals with a simpler model (`ASteRA`) that neglects this in order to explore under which conditions this additional complexity is necessary. We find that `ASteRA` performs very well in cases of weak to moderate stellar contamination, constraining the planetary parameters to a high

degree of accuracy. Importantly, in all cases, ASteRA performs far better than no correction attempt at all, consistent with the findings of previous studies. The need for an activity correction is especially demonstrated by the retrieved H_2O mixing ratios, which are underestimated by over two orders of magnitude in the worst-case scenario when no correction is applied. Iyer & Line (2020) find that stellar contamination must be accounted for when spot-filling factors exceed 1% in order to avoid biasing the retrieved planetary parameters. Our results substantiate this, however, we highlight that the spot temperature contrast is also important to consider alongside the filling factor. The Bayes factor is still moderately ($lnB = 4.12$) to strongly ($lnB = 11.33$) in favor of the model containing the activity correction in the cases of a small ($0.1R_*$), high-contrast ($\Delta T_{\rm spot}$= 1000 K) spot at latitudes of 30° and 0°, respectively (Cases 1 and 2), despite both of these spots having filling factors $\leqslant$1%. The spot latitude also contributes but to a lesser degree than the temperature contrast in the small spot regime. Importantly, there is no evidence for a hard boundary or on–off switch for where stellar contamination should be considered. It will therefore be beneficial to apply an activity correction to safeguard against bias even in the lower-activity regimes where contamination is less dominant.

Degeneracies between spot parameters at low resolution mean that the retrievals tend to be less successful in accurately characterizing the host star, particularly in regimes of low-to-moderate activity. Nevertheless, from a planetary perspective, it is an excellent tool for accounting for potential contamination bias in the fundamental planetary properties R_P, T_P, and $\log(H_2O)$. We believe that this is how it should be viewed and utilized by the community moving forward, albeit with the caveat that it cannot remove any bias that is introduced as a result of the inaccurate characterization of fundamental stellar parameters.

For the highest-activity cases considered here, the bias introduced by stellar contamination cannot be fully corrected for by ASteRA due to the limb darkening–spot interplay being neglected. As such, in these scenarios, a small amount of residual bias remains resulting in a slight loss of accuracy in the retrieved planetary parameters. This subtle loss of accuracy would not be easily identified without a priori knowledge of the activity level of the star. For this reason, it may be beneficial to consider other stellar activity mitigation processes in parallel if these are available, e.g., utilizing the out-of-transit observations or continuous photometry, in order to better interpret the retrieval results in the context of the host star. We show that for multiple-spot cases the bias introduced by the limb darkening–spot interplay is minimized under the assumption that all spots have the same temperature. As such, single-large spots present the worst-case scenario for real observations. Further bias is also introduced if these large spots have separated into umbral and penumbral regions characterized by different distinct temperatures, making this something we should progressively start to consider as a community.

Although the results of this study are based on idealized spectra, the spot cases investigated in this analysis will provide a good baseline from which to fully explore the impact of stellar activity on both JWST and Ariel observations as our simulations cover a similar wavelength range and spectral resolution to what is/will be obtainable with these observatories. In future work, we intend to conduct similar investigations using more realistic models and observations. With this in mind, from a stellar perspective, we aim to extend the flexibility of the ASteRA plugin to incorporate the interplay of spots and limb darkening, as this study has shown that there are scenarios where this cannot reasonably be neglected. This will, however, need to be done with caution to avoid unknowingly injecting bias. We also aim to extend our retrievals to investigate other spectral types, in particular, M dwarfs, which may be more complicated, especially as their spectra can contain molecular lines that could be incorrectly attributed to the exoplanet atmosphere. Other exciting lines of investigation that naturally follow from this work include exploring more complex manifestations of stellar activity, for example, occulted spots and the presence of both spots and faculae. With respect to the exoplanetary atmosphere, we intend to extend this analysis to more complex, realistic atmospheric compositions. A particular emphasis will be put on physical processes responsible for producing features in the optical regime, where the stellar contamination is most pronounced, such as the presence of clouds and hazes and other opacity sources, e.g., absorption by the alkali metals Na and K. Finally, we intend to transition from using idealized spectra to simulated instrument observations to explore the effects of more realistic noise and a more restricted wavelength coverage in the optical before eventually using this framework to achieve our ultimate goal of accurately analyzing real observations.

Acknowledgments

The authors would like to thank the anonymous referee for the insightful comments. A.T. would like to thank Dr Kai Hou Yip for the insightful discussions. The work presented in this paper was partially supported by UKSA, grant ST/X002616/1. The authors also acknowledge the support of the ARIEL ASI-INAF agreement n.2021-5-HH.0. G.M. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 895525, and from the Ariel Postdoctoral Fellowship program of the Swedish National Space Agency (SNSA). This work would not have been possible without the use of publicly available data and open-source software, as such the authors also wish to reiterate their acknowledgment of the use of TauREx3 (Al-Refaie et al. 2021) https://github.com/ucl-exoplanets/TauREx3_public, ExoTETHyS (Morello et al. 2020a, 2020b, 2021) https://github.com/ucl-exoplanets/ExoTETHyS, pylightcurve (Tsiaras et al. 2016) https://github.com/ucl-exoplanets/pylightcurve, the PHOENIX BT-Settl Library (Allard et al. 2012; Baraffe et al. 2015), MultiNest (Feroz et al. 2009), and PyMultiNest (Buchner et al. 2014).

Appendix A
Posterior Distributions Retrieved with ASteRA for Cases 7–9 (with and without Limb Darkening)

This Appendix contains the posterior distributions for the single-spot cases 7–9 investigated in Section 4.4 and displayed in Figures 13–15, respectively.

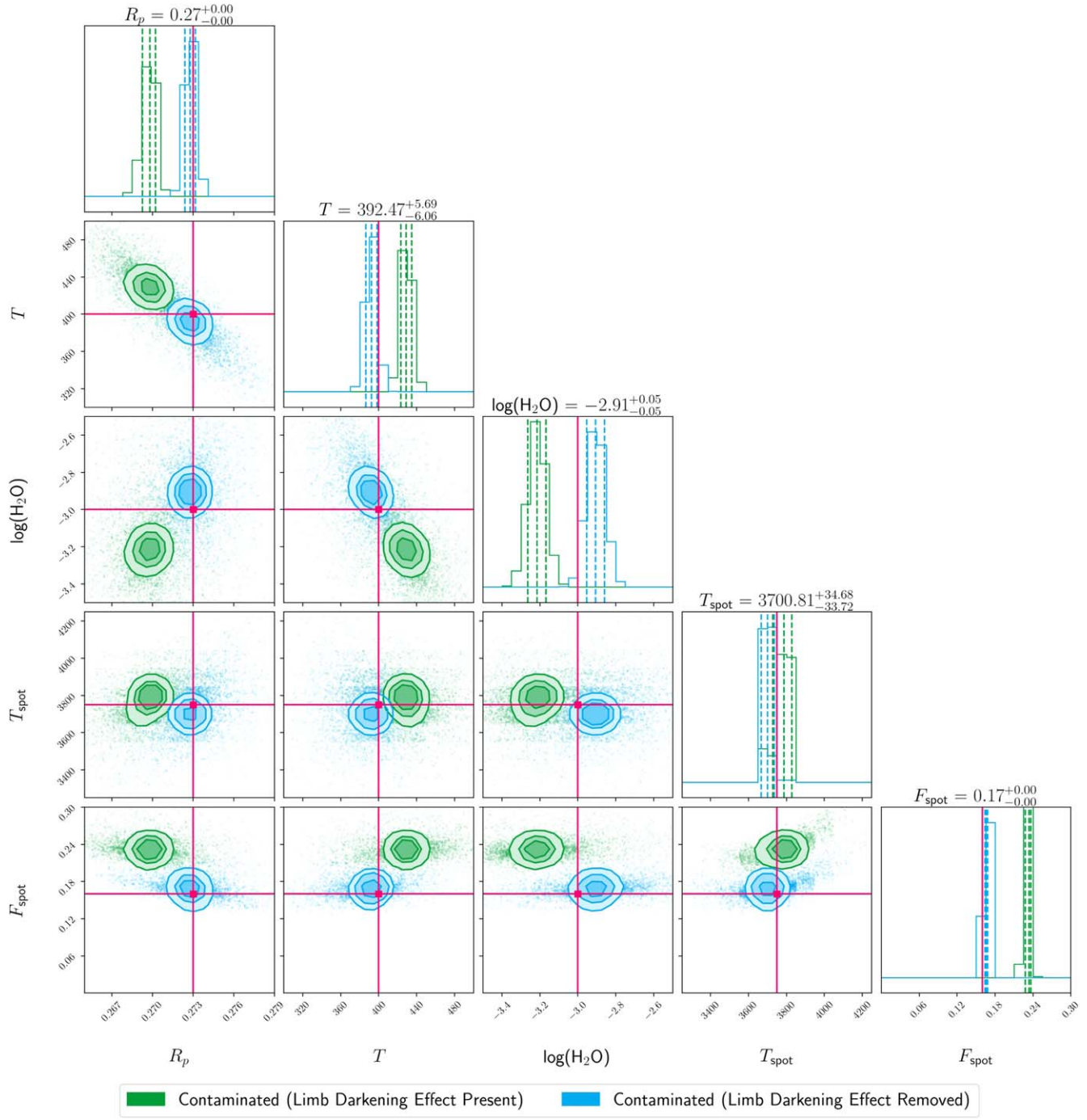

Figure 13. Retrieved posterior distributions for spot Case 7 when retrievals were conducted on a spot-contaminated spectrum, including contributions from the interplay of the spot and the limb-darkening effect (green) and for a spot-contaminated spectrum when the limb-darkening contribution is removed i.e., the LDCs are set to zero (blue). The retrieved values given for each parameter above each column correspond to the blue posteriors. The ground truth values for each parameter are indicated by the pink lines.

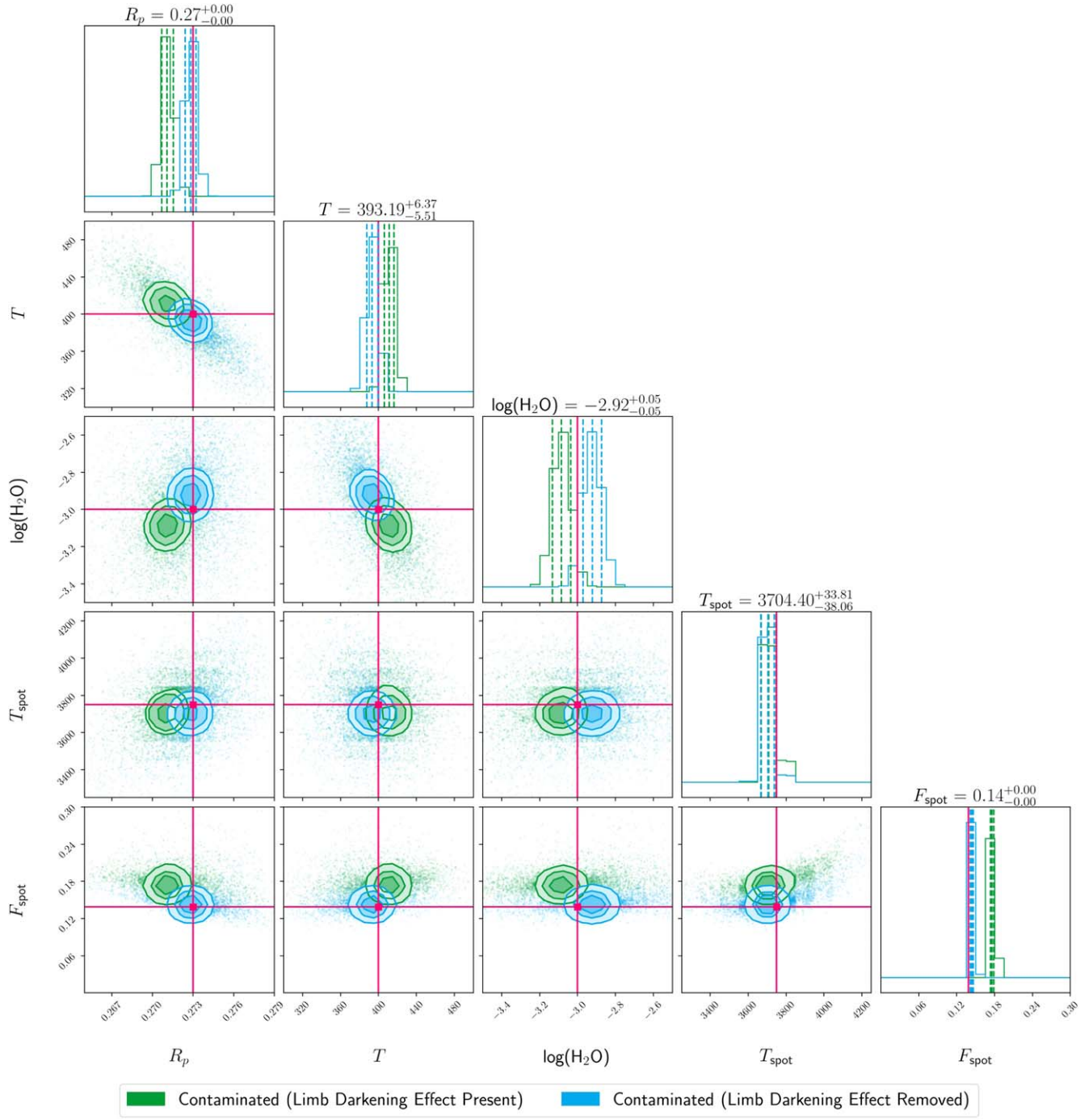

Figure 14. Retrieved posterior distributions for spot Case 8. Figure elements are the same as those in Figure 13.

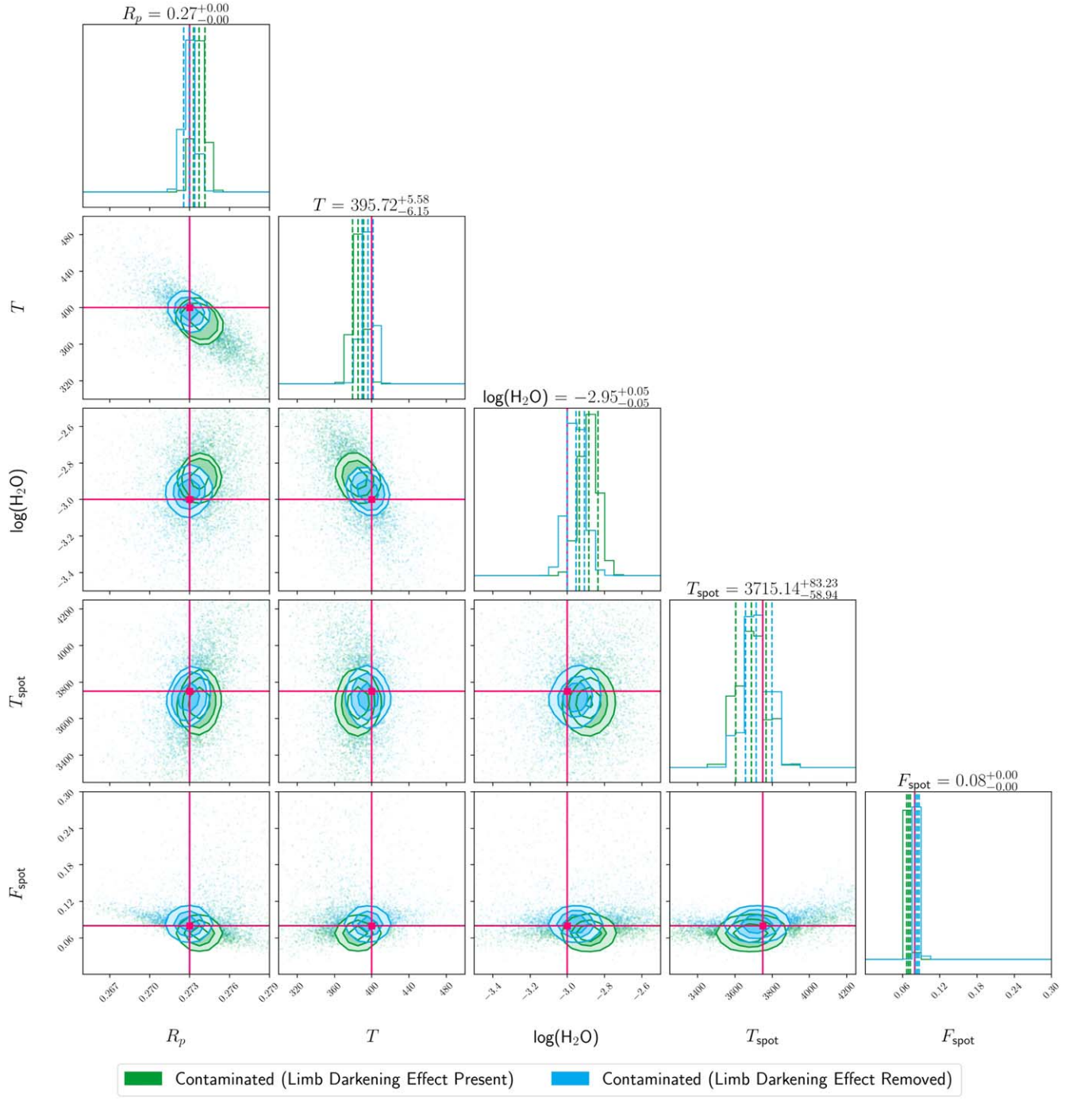

Figure 15. Retrieved posterior distributions for spot Case 9. Figure elements are the same as those in Figure 13.

Appendix B
Posterior Distribution for the Multiple-spot Cases

This Appendix contains the posterior distributions for the multiple-spot cases tested in Section 4.5. These are the two-spot case (Figure 16), the multiple-spots with random configuration case (Figure 17), and the multiple-spots with a preferred active latitude case (Figure 18), respectively.

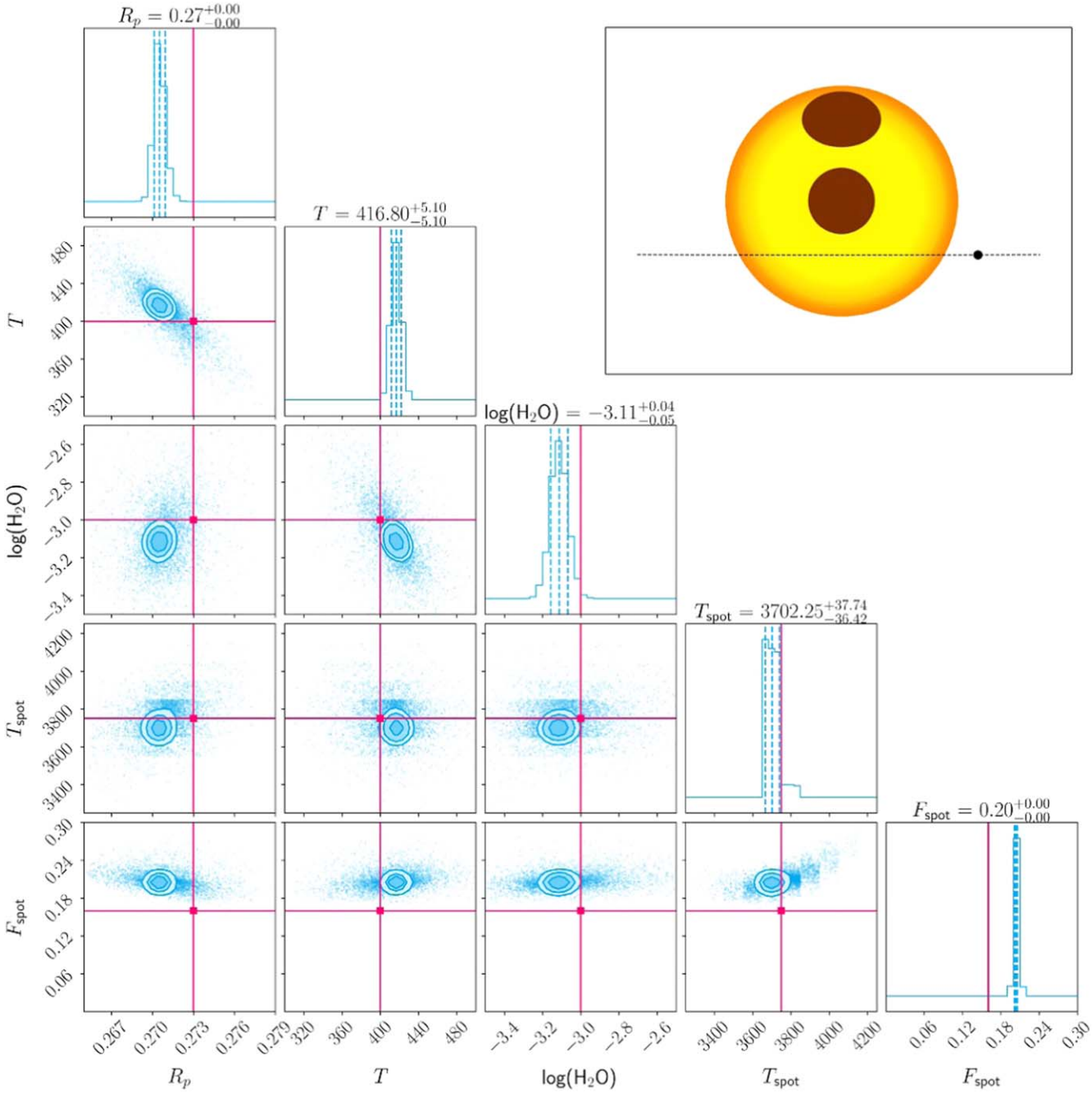

Figure 16. Retrieved posterior distributions for the two-spot case (Section 4.5). The pink lines indicate the ground truth values for each parameter. Inset: a graphical depiction of the stellar disk.

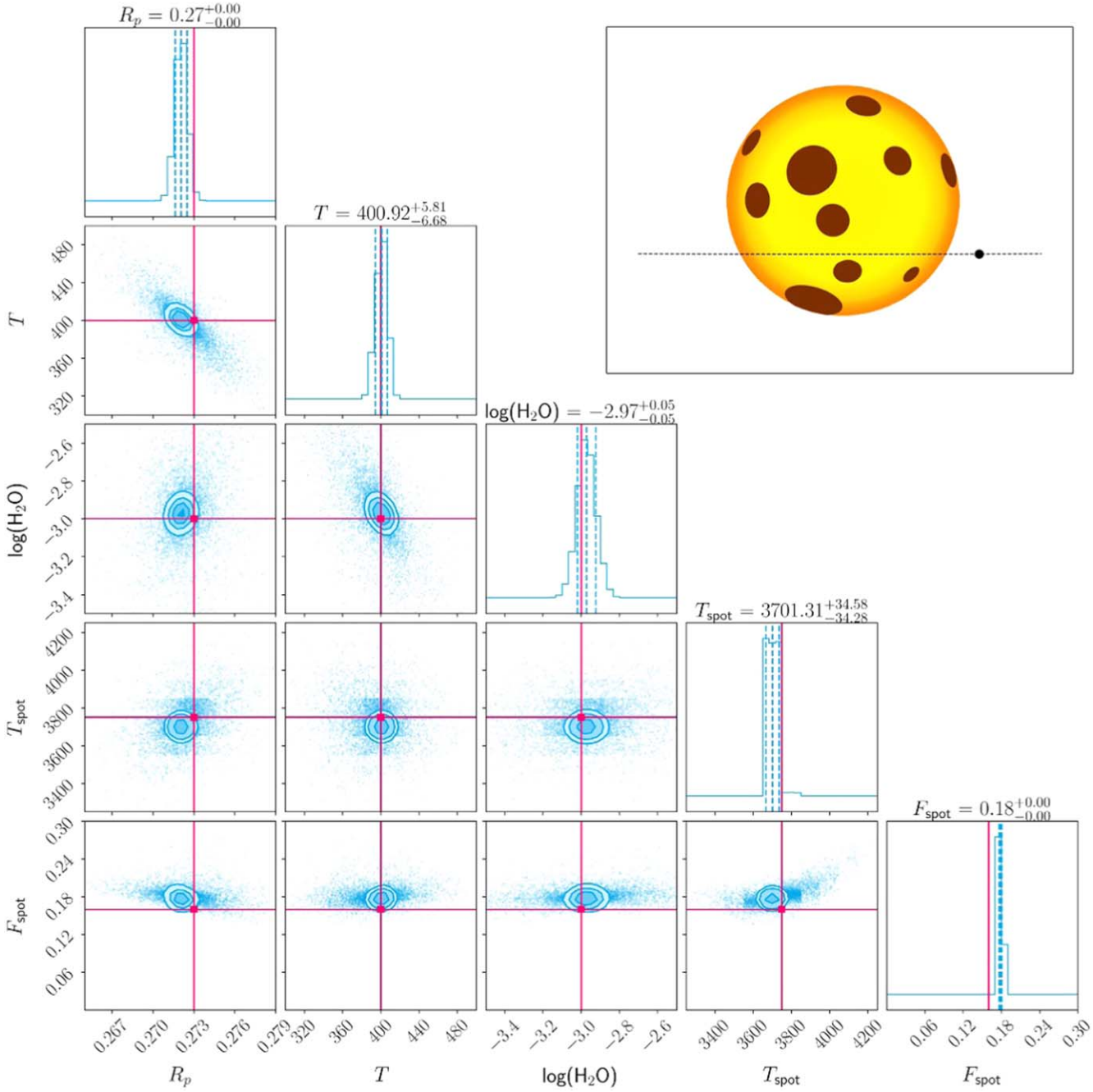

Figure 17. Retrieved posterior distributions for the multiple-spot case in which the spots have a random configuration (Section 4.5). The pink lines indicate the ground truth values for each parameter. Inset: a graphical depiction of the stellar disk.

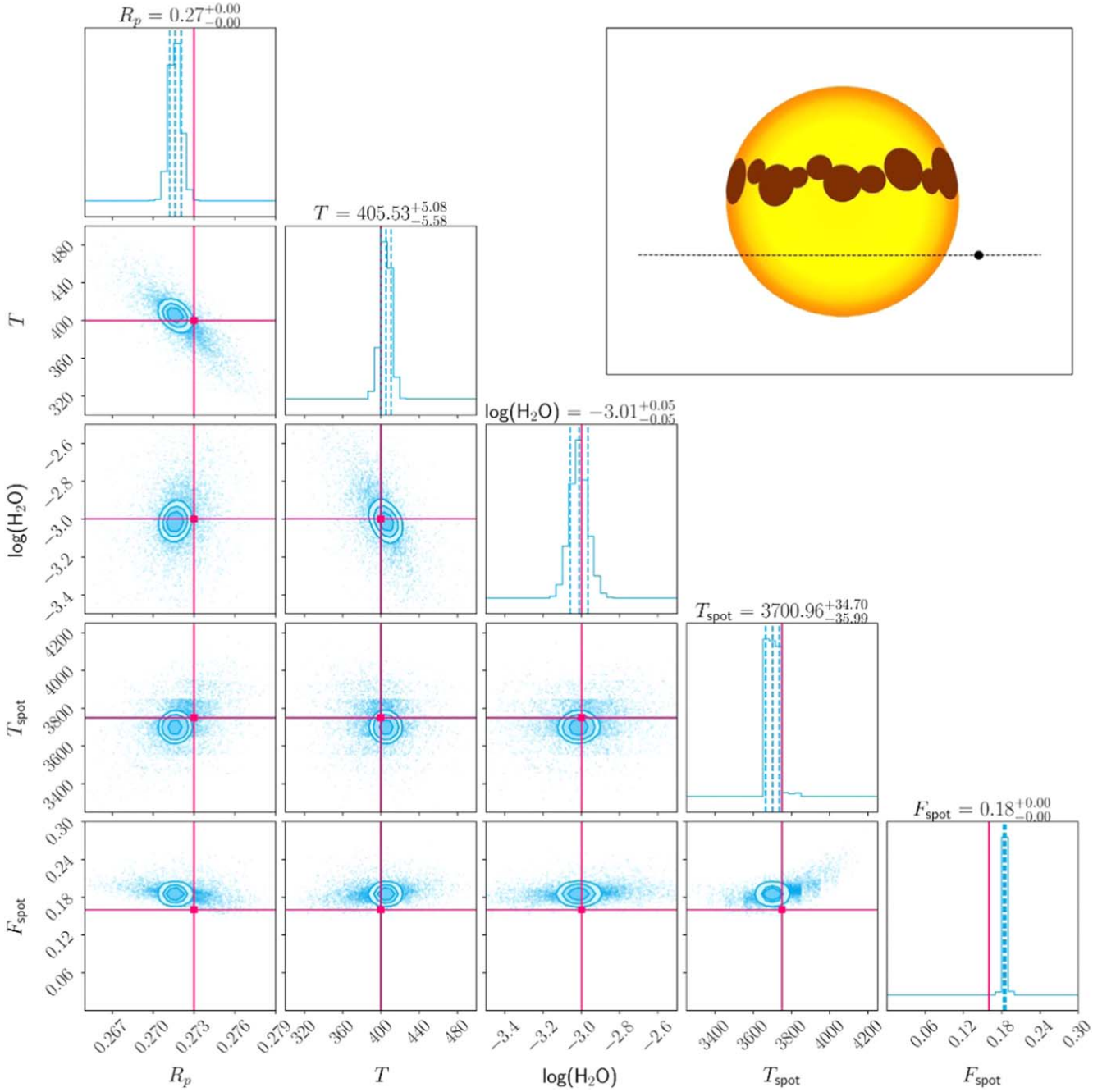

Figure 18. Retrieved posterior distributions for the multiple-spot case in which the spots occupy a preferred active latitude centered around $\phi = 15°$ (Section 4.5). The pink lines indicate the ground truth values for each parameter. Inset: a graphical depiction of the stellar disk.

Appendix C
Posterior Distributions for the Radial Temperature Variation Spot Cases

This Appendix contains the posterior distributions for the cases with spots displaying radial temperature variations, i.e., a separation into umbra and penumbra, as discussed in Section 4.6. The posteriors given correspond to the radial temperature variation Case 1 (Figure 19) and the radial temperature variation Case 2 (Figure 20), respectively.

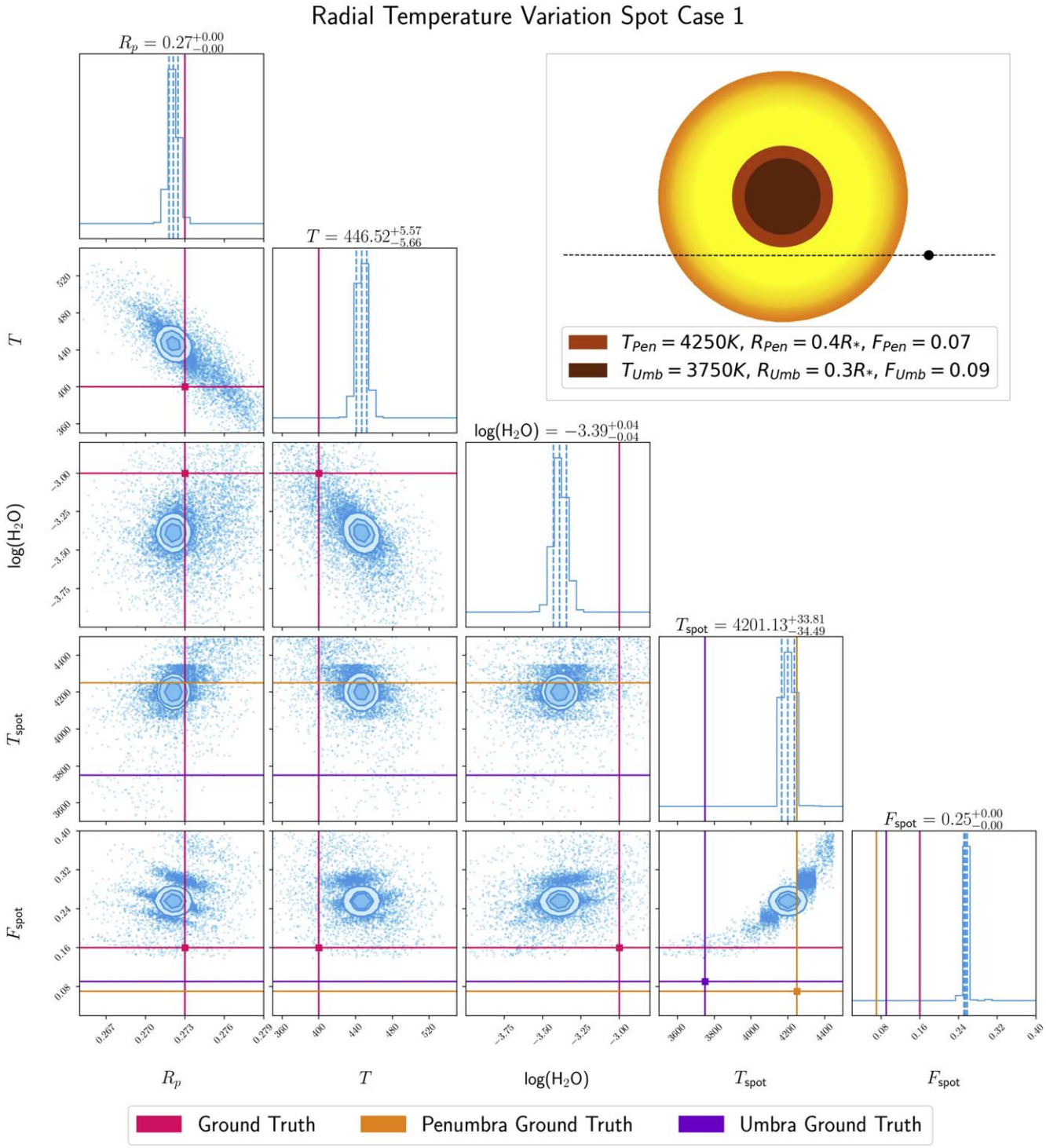

Figure 19. Retrieved posterior distributions for Case 1 when considering a spot with radial temperature variations (Section 4.6). The pink lines indicate the ground truth values for each parameter. The orange and purple lines depict the ground truth values for the spot penumbra and umbra, respectively. Inset: a graphical depiction of the stellar disk for this case showing the separation of the spot into an umbra and penumbra and their respective parameters.

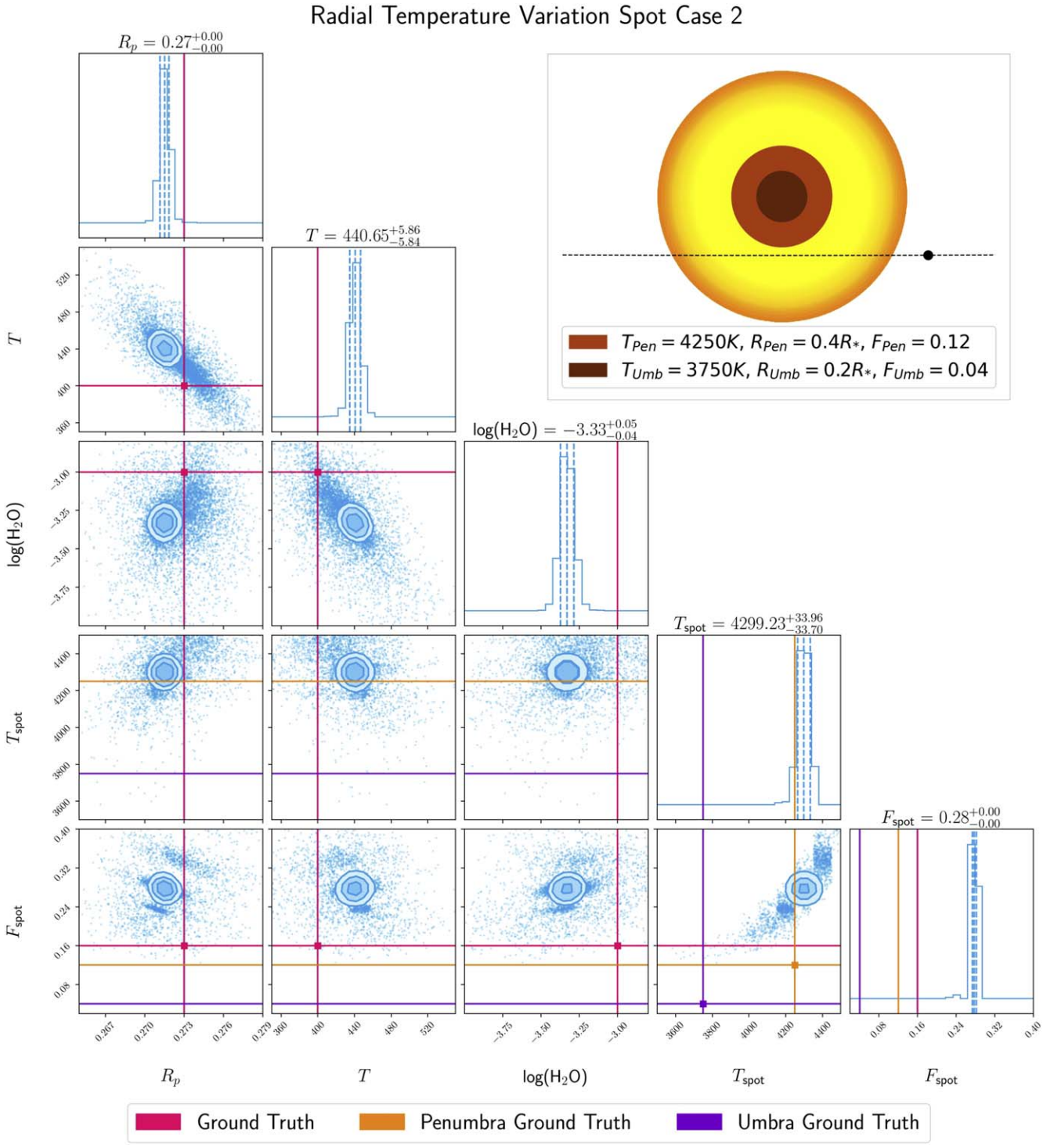

Figure 20. Retrieved posterior distributions for Case 2 when considering a spot with radial temperature variations (Section 4.6). Figure elements are the same as those in Figure 19.

ORCID iDs

Alexandra Thompson https://orcid.org/0000-0003-4128-2270
Antonino Petralia https://orcid.org/0000-0002-9882-1020
Quentin Changeat https://orcid.org/0000-0001-6516-4493
Arianna Saba https://orcid.org/0000-0002-1437-4228
Giuseppe Morello https://orcid.org/0000-0002-4262-5661
Mario Morvan https://orcid.org/0000-0001-8587-2112
Giuseppina Micela https://orcid.org/0000-0002-9900-4751
Giovanna Tinetti https://orcid.org/0000-0001-6058-6654

References

Al-Refaie, A. F., Changeat, Q., Waldmann, I. P., et al. 2021, ApJ, 917, 37
Allard, F., Homeier, D., & Freytag, B. 2012, RSPTA, 370, 2765
Almenara, J. M., Bonfils, X., Forveille, T., et al. 2022, A&A, 667, L11
Al-Refaie, A. F., Changeat, Q., Venot, O., et al. 2022, ApJ, 932, 123
Anisman, L. O., Edwards, B., Changeat, Q., et al. 2020, AJ, 160, 233

Arcangeli, J., Désert, J.-M., Parmentier, V., et al. 2019, A&A, 625, A136
Ballerini, P., Micela, G., Lanza, A. F., et al. 2012, A&A, 539, A140
Baraffe, I., Homeier, D., Allard, F., et al. 2015, A&A, 577, A42
Bean, J. L., Stevenson, K. B., Batalha, N. M., et al. 2018, PASP, 130, 114402
Berdyugina, S. V. 2005, LRSP, 2, 8
Bixel, A., Rackham, B. V., Apai, D., et al. 2019, AJ, 157, 68
Bradshaw, S. J., & Hartigan, P. 2014, ApJ, 795, 79
Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125
Cauley, P. W., Kuckein, C., Redfield, S., et al. 2018, AJ, 156, 189
Changeat, Q. 2022, AJ, 163, 106
Changeat, Q., Edwards, B., Al-Refaie, A. F., et al. 2022, ApJS, 260, 3
Charbonneau, D., Brown, T. M., Noyes, R. W., et al. 2002, ApJ, 568, 377
Ciardi, D. R., von Braun, K., Bryden, G., et al. 2011, AJ, 141, 108
Claret, A. 2000, A&A, 363, 1081
Claret, A., Hauschildt, P. H., & Witte, S. 2012, A&A, 546, A14
Claret, A., Hauschildt, P. H., & Witte, S. 2013, A&A, 552, A16
Cracchiolo, G., Micela, G., Morello, G., et al. 2021a, MNRAS, 507, 6118
Cracchiolo, G., Micela, G., & Peres, G. 2021b, MNRAS, 501, 1733
Crouzet, N., McCullough, P. R., Deming, D., et al. 2014, ApJ, 795, 166
Czesla, S., Huber, K. F., Wolter, U., et al. 2009, A&A, 505, 1277
Dang, L., Bell, T. J., Cowan, N. B., et al. 2022, AJ, 163, 32
Danielski, C., Brucalassi, A., Benatti, S., et al. 2022, ExA, 53, 473
Dressing, C. D., & Charbonneau, D. 2015, ApJ, 807, 45
Edwards, B., Changeat, Q., Tsiaras, A., et al. 2023, ApJS, 269, 31
Edwards, B., Changeat, Q., Mori, M., et al. 2021, AJ, 161, 44
Espinoza, N., Rackham, B. V., Jordán, A., et al. 2019, MNRAS, 482, 2065
Evans, T. M., Sing, D. K., Kataria, T., et al. 2017, Natur, 548, 58
Feinstein, A. D., France, K., Youngblood, A., et al. 2022, AJ, 164, 110
Feng, Y. K., Line, M. R., & Fortney, J. J. 2020, AJ, 160, 137
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Garcia, R. A., Ballot, J., Mathur, S., et al. 2010, arXiv:1012.0494
Gomes da Silva, J., Santos, N. C., Bonfils, X., et al. 2011, A&A, 534, A30
Goulding, N. T., Barnes, J. R., Pinfield, D. J., et al. 2012, MNRAS, 427, 3358
Gressier, A., Mori, M., Changeat, Q., et al. 2022, A&A, 658, A133
Gully-Santiago, M. A., Herczeg, G. J., Czekala, I., et al. 2017, ApJ, 836, 200
Hartman, J. D., Bakos, G. Á., Noyes, R. W., et al. 2011, AJ, 141, 166
Herrero, E., Ribas, I., Jordi, C., et al. 2016, A&A, 586, A131
Hoeijmakers, H. J., Ehrenreich, D., Kitzmann, D., et al. 2019, A&A, 627, A165
Howarth, I. D. 2011, MNRAS, 418, 1165
Husser, T.-O., Wende-von Berg, S., Dreizler, S., et al. 2013, A&A, 553, A6
Irwin, P. G. J., Parmentier, V., Taylor, J., et al. 2020, MNRAS, 493, 106
Iyer, A. R., & Line, M. R. 2020, ApJ, 889, 78
Jackson, R. J., & Jeffries, R. D. 2013, MNRAS, 431, 1883
Järvinen, S. P., Strassmeier, K. G., Carroll, T. A., et al. 2018, A&A, 620, A162
Klein, B., Zicher, N., Kavanagh, R. D., et al. 2022, MNRAS, 512, 5067
Knutson, H. A., Lewis, N., Fortney, J. J., et al. 2012, ApJ, 754, 22
Kokori, A., Tsiaras, A., Edwards, B., et al. 2022, ExA, 53, 547
Korolik, M., Roettenbacher, R. M., Fischer, D. A., et al. 2023, AJ, 166, 123
Maggio, A., Locci, D., Pillitteri, I., et al. 2022, ApJ, 925, 172
Magrini, L., Danielski, C., Bossini, D., et al. 2022, A&A, 663, A161
Mayor, M., Pepe, F., Queloz, D., et al. 2003, Msngr, 114, 20
McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24
Micela, G. 2015, ExA, 40, 723
Mikal-Evans, T., Sing, D. K., Kataria, T., et al. 2020, MNRAS, 496, 1638
Morello, G., Claret, A., Martin-Lagarde, M., et al. 2020a, AJ, 159, 75
Morello, G., Claret, A., Martin-Lagarde, M., et al. 2020b, JOSS, 5, 46
Morello, G., Zingales, T., Martin-Lagarde, M., et al. 2021, AJ, 161, 174
Morris, B. M. 2020, ApJ, 893, 67
Morris, B. M., Hebb, L., Davenport, J. R. A., et al. 2017, ApJ, 846, 99
Newton, E. R., Irwin, J., Charbonneau, D., et al. 2016, ApJL, 821, L19
Öberg, K. I., Murray-Clay, R., & Bergin, E. A. 2011, ApJL, 743, L16
Oshagh, M., Santos, N. C., Ehrenreich, D., et al. 2014, A&A, 568, A99
Pacetti, E., Turrini, D., Schisano, E., et al. 2022, ApJ, 937, 36
Panja, M., Cameron, R., & Solanki, S. K. 2020, ApJ, 893, 113
Patel, J. A., & Espinoza, N. 2022, AJ, 163, 228
Pinhas, A., Madhusudhan, N., Gandhi, S., et al. 2019, MNRAS, 482, 1485
Pinhas, A., Rackham, B. V., Madhusudhan, N., et al. 2018, MNRAS, 480, 5314
Pluriel, W., Whiteford, N., Edwards, B., et al. 2020, AJ, 160, 112
Pont, F., Sing, D. K., Gibson, N. P., et al. 2013, MNRAS, 432, 2917
Rackham, B. V., Apai, D., & Giampapa, M. S. 2018, ApJ, 853, 122
Rackham, B. V., Apai, D., & Giampapa, M. S. 2019, AJ, 157, 96
Raymond, S. N., & Morbidelli, A. 2022, in Demographics of Exoplanetary Systems, ed. K. Biazzo, V. Bozza, L. Mancini, & A. Sozzetti, Vol. 466 (Berlin: Springer), 3
Reiners, A., & Basri, G. 2010, ApJ, 710, 924
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, Proc. SPIE, 9143, 914320
Rustamkulov, Z., Sing, D. K., Mukherjee, S., et al. 2023, Natur, 614, 659
Saba, A., Tsiaras, A., Morvan, M., et al. 2022, AJ, 164, 2
Sanchis-Ojeda, R., & Winn, J. N. 2011, ApJ, 743, 61
Silva-Valio, A., Lanza, A. F., Alonso, R., et al. 2010, A&A, 510, A25
Sing, D. K., Fortney, J. J., Nikolov, N., et al. 2016, Natur, 529, 59
Sing, D. K., Pont, F., Aigrain, S., et al. 2011, MNRAS, 416, 1443
Skaf, N., Bieger, M. F., Edwards, B., et al. 2020, AJ, 160, 109
Solanki, S. K. 2003, A&ARv, 11, 153
Stevenson, K. B., Désert, J.-M., Line, M. R., et al. 2014, Sci, 346, 838
Strassmeier, K. G., Carroll, T. A., & Ilyin, I. V. 2023, A&A, 674, A118
Swain, M. R., Vasisht, G., Tinetti, G., et al. 2008, Natur, 452, 329
Szabó, G. M., Gandolfi, D., Brandeker, A., et al. 2021, A&A, 654, A159
Thao, P. C., Mann, A. W., Gao, P., et al. 2023, AJ, 165, 23
Tinetti, G., Eccleston, P., Haswell, C., et al. 2021, arXiv:2104.04824
Tinetti, G., Vidal-Madjar, A., Liang, M.-C., et al. 2007, Natur, 448, 169
Trotta, R. 2008, ConPh, 49, 71
Tsiaras, A., Waldmann, I. P., Rocchetto, M., et al., 2016 pylightcurve: Exoplanet lightcurve model, Astrophysics Source Code Library, ascl:1612.018
Tsiaras, A., Waldmann, I. P., Tinetti, G., et al. 2019, NatAs, 3, 1086
Tsiaras, A., Waldmann, I. P., Zingales, T., et al. 2018, AJ, 155, 156
Venot, O., Parmentier, V., Blecic, J., et al. 2020, ApJ, 890, 176
von Essen, C., Mallonn, M., Borre, C. C., et al. 2020, A&A, 639, A34
Yee, S. W., Petigura, E. A., Fulton, B. J., et al. 2018, AJ, 155, 255
Yip, K., Changeat, Q., Al-Refaie, A., et al. 2022, arXiv:2205.07037
Zellem, R. T., Swain, M. R., Roudier, G., et al. 2017, ApJ, 844, 27
Zhang, Z., Zhou, Y., Rackham, B. V., et al. 2018, AJ, 156, 178