Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **512,** 2604–2617 (2022) https://doi.org/10.1093/mnras/stac662
Advance Access publication 2022 March 12

Applications of a Gaussian process framework for modelling of high-resolution exoplanet spectra

Annabella Meech [1]★ Suzanne Aigrain,[1] Matteo Brogi[2] and Jayne L. Birkby [1]

[1]*Physics Department, University of Oxford, Denys Wilkinson Building, Oxford OX1 3RH, UK*
[2]*Physics Department, University of Warwick, Coventry CV4 7AL, UK*

Accepted 2022 March 2. Received 2022 February 18; in original form 2021 October 12

ABSTRACT

Observations of exoplanet atmospheres in high resolution have the potential to resolve individual planetary absorption lines, despite the issues associated with ground-based observations. The removal of contaminating stellar and telluric absorption features is one of the most sensitive steps required to reveal the planetary spectrum and, while many different detrending methods exist, it remains difficult to directly compare the performance and efficacy of these methods. Additionally, though the standard cross-correlation method enables robust detection of specific atmospheric species, it only probes for features that are expected a priori. Here, we present a novel methodology using Gaussian process (GP) regression to directly model the components of high-resolution spectra, which partially addresses these issues. We use two archival CRyogenic Infra-Red Echelle Spectrograph (CRIRES)/Very Large Telescope (VLT) data sets as test cases, observations of the hot Jupiters HD 189733 b and 51 Pegasi b, recovering injected signals with average line contrast ratios of $\sim 4.37 \times 10^{-3}$ and $\sim 1.39 \times 10^{-3}$, and planet radial velocities $\Delta K_{\rm p} = 1.45 \pm 1.53\,{\rm km\,s^{-1}}$ and $\Delta K_{\rm p} = 0.12 \pm 0.12\,{\rm km\,s^{-1}}$ from the injection velocities, respectively. In addition, we demonstrate an application of the GP method to assess the impact of the detrending process on the planetary spectrum, by implementing injection-recovery tests. We show that standard detrending methods used in the literature negatively affect the amplitudes of absorption features in particular, which has the potential to render retrieval analyses inaccurate. Finally, we discuss possible limiting factors for the non-detections using this method, likely to be remedied by higher signal-to-noise data.

Key words: atmospheric effects – methods: data analysis – techniques: spectroscopic – planets and satellites: atmospheres – infrared: planetary systems.

1 INTRODUCTION

High-resolution spectroscopy (HRS) is one of the current leading techniques used to probe exoplanet atmospheres and, since its conception (Snellen et al. 2010), has enabled inference of the presence of atomic and molecular species, atmospheric dynamics, and temperature structure (Brogi et al. 2012; Rodler, Lopez-Morales & Ribas 2012; Birkby et al. 2017; Nugroho et al. 2017; Flowers et al. 2019; Ehrenreich et al. 2020; Pino et al. 2020; Beltz et al. 2021; Wardenier et al. 2021). From the Earth, we observe the composite spectrum

$$F_{\rm obs} = (F_* + F_{\rm p}) \times T \times A, \tag{1}$$

where F_*, $F_{\rm p}$, and T are the stellar, planet, and telluric spectra, respectively, altered by the instrumental transmission A, with each component a function of time and wavelength. F_* and $F_{\rm p}$ shift according to the stellar reflex motion and planet orbital motion, respectively, the latter typically orders of magnitude greater. The principal hurdle is distinguishing the relatively weak planet signal from the multitude of nuisance signals, the dominant contributions being the stellar and variable telluric absorption. Though occasionally individual planet atmospheric absorption lines can be distinguished (Brogi et al. 2012; Schwarz et al. 2016), they typically have a signal-to-noise ratio (SNR) smaller than 1, even when combining multiple nights of data. The key insight is that the planet's rest frame moves significantly with respect to the other components given in equation (1). Therefore, if observed at different orbital phases, we can isolate the exoplanet's transmission or emission spectrum by its Doppler shift.

One of the prevailing challenges to maximizing HRS capabilities is the effective removal of telluric contamination, while preserving the true astronomical signal. Though telluric signals are considered stationary in wavelength, the rapid fluctuations of the Earth's atmospheric conditions cause variation in depth and shape over short time-scales. Observing bands in the near-infrared (NIR) offers a more favourable planet-to-star contrast ratio; however, there are regions of strong telluric absorption at these wavelengths due to dominant species CH_4, CO_2, and H_2O. Many different telluric detrending techniques have been successfully implemented for HRS. Theoretical atmospheric transmission modelling, with codes such as TELFIT and MOLECFIT (Gullikson, Dodson-Robinson & Kraus 2014; Smette et al. 2015), combines a radiative transfer treatment of the Earth's atmosphere with observatory metadata to build a synthetic telluric model. Though very effective in the optical (Casasayas-Barris et al. 2019; Bourrier et al. 2020; Hoeijmakers et al. 2020), such modelling fails to capture variations in the heavy absorption seen at longer wavelengths; thus, empirical methods have been favoured in NIR

★ E-mail: annabella.meech@physics.ox.ac.uk

studies. Some have opted for principal component analysis (PCA) to remove contaminating features: De Kok et al. (2013) were one of the first to apply this decomposition of the spectral matrix in order to identify static telluric features. Later, Piskorz et al. (2016, 2017, 2018) applied a similar technique to NIRSPEC/Keck observations, guiding the PCA with a baseline telluric model. The recent study of HD 209508 b then further developed the PCA technique to detrend GIANO-B/TNG spectra (Giacobbe et al. 2021). SYSREM, a specific, data point uncertainty-weighted PCA-based algorithm (Tamuz, Mazeh & Zucker 2005; Mazeh, Tamuz & Zucker 2007), has been widely used to correct for common mode systematic effects in high-resolution spectral data, and has proved to be a powerful tool, particularly for removal of strong contamination (Birkby et al. 2013; Nugroho et al. 2017; Gibson et al. 2020; Kesseli et al. 2020; Merritt et al. 2020). For each pass, the SYSREM algorithm produces a low-rank approximation of the spectral matrix, a product of two column vectors, identifying common modes in each wavelength channel that can then be subtracted (see Tamuz et al. 2005 for further detail). The low-order trends are often correlated with physical conditions such as airmass and seeing (Birkby et al. 2017). SYSREM requires fine-tuning to carefully remove as much of the telluric contamination as possible without infringing on excessive use, detrimental to the planet spectrum. Another popular technique uses linear regression to model the variation of flux with airmass, combined with additional sampling of particular wavelengths known to accommodate strong telluric lines (Brogi et al. 2012, 2013, 2014; Webb et al. 2020). In this method, the fluxes in each wavelength channel (pixel) are divided through by a least-squares fitted linear trend with airmass. This first pass captures the low-order trends of the telluric variation. Then, the evolution of residuals in each individual pixel is modelled linearly with the temporal variation of the (summed) residuals at locations of known strong telluric lines. Finally, a high-pass filter is applied, involving outlier rejection and removal of any remaining trends in the spectral direction. While SYSREM searches for any common modes in wavelength, including any quasi-stationary stellar lines, the assumption of airmass variation in the latter method makes it unsuitable when strong stellar features are present.

Given the plethora of techniques that can be used to remove telluric contamination, it can be difficult to choose the most suitable approach for a particular data set, considering the atmospheric conditions on the night of observation, and spectral range. Some studies have used the cross-correlation detection significance to compare different detrending algorithms (Cabot et al. 2019; Langeveld et al. 2021). Moreover, for atmospheric characterization, we are concerned with the impact on the planet spectrum. It is widely accepted that most detrending algorithms alter the planet signal in some way; hence, it is important to subject the cross-correlation models to the same processing for accurate retrieval work (Brogi & Line 2019; Gibson et al. 2020). However, since the planet spectrum goes unseen with these standard methods, it is not straightforward to pinpoint the modifications.

High-resolution cross-correlation spectroscopy (HRCCS) enables detection of species by way of a line-matching exercise: the residual, corrected fluxes are cross-correlated with a template planetary spectrum, derived by assuming a radiative transfer treatment, temperature profile, and abundance of chemical species in the planet's atmosphere. The planet signal is typically much weaker than the noise of the detrended fluxes, and even the photon noise level. Since the SNR of the planet is proportional to the number of detected lines (Birkby 2018), HRCCS combines the signal from each line: the cross-correlation functions (CCFs) resulting from each spectrum are aligned into the assumed planet rest frame and summed. This is repeated for a range of systemic and planet radial velocities. Then, there are various metrics used to compute the detection significance from the cross-correlation velocity maps, notably comparing the signal of the cross-correlation peak to the standard deviation of 'noise' outside the peak, and separately the Welch *t*-test, whereby the null hypothesis that the CCFs within the planet trail are drawn from the same parent distribution as those outside the trail is rejected at a certain significance (Birkby et al. 2013, 2017; Brogi et al. 2013; De Kok et al. 2013). Both methods have their known flaws when it comes to deriving reliable detection significances (Cabot et al. 2019) and translating to estimates of atmospheric properties. There has been work to map the cross-correlation values to a log-likelihood, in order to enable integration into a Bayesian analysis retrieval framework (Brogi & Line 2019; Gibson et al. 2020), akin to low-resolution spectroscopy retrievals (Madhusudhan & Seager 2009). These approaches then allow more statistically robust model testing by exploration of the full parameter space, and have higher constraining power. That said, they assume that the data are independent, and the likelihood of Gibson et al. (2020) assumes that all data point uncertainties are uniformly scaled. This may be problematic in some cases, for example when dealing with both regions of continuum and cores of telluric lines.

In this work, we present a Gaussian process (GP) regression methodology to directly model the telluric and planet signals sequentially, analyse its potential, and compare its efficacy to the results of current popular approaches. In Section 2, we present the GP framework used to forward model the composite spectrum. We investigate the performance of sequential GP modelling by application to archival NIR observations of the dayside of two well-studied hot Jupiters, HD 189733 b and 51 Peg b; the data reduction and results of these tests are described in Section 3. We then discuss an application of the method, namely the comparison of different telluric corrections in Section 4, and our conclusions are outlined in Section 5.

2 PROPOSED GP MODELLING OF HIGH-RESOLUTION SPECTRA

There have been many successful examples of applying GPs to time-series exoplanet data, modelling stellar variability for exoplanet radial velocity extraction, stellar-rotation periods, photometry, and systematics for (e.g.) low-resolution transmission spectroscopy (Gibson et al. 2012; Rajpaul et al. 2015; Aigrain, Parviainen & Pope 2016; Angus et al. 2018). Czekala et al. (2017) first presented the concept of employing GPs to disentangle different components in high-resolution spectra, specifically binary stellar spectra, with the package PSOAP. Their method was limited by computational expense, only reasonably allowing evaluation of the GP on narrow bandpasses and requiring access to computer clusters. We propose that GP regression offers an attractive route for direct inference of the components of exoplanet high-resolution spectra in a principled, probabilistic manner. Here, we attempt to create a reasonably fast GP framework for this purpose.

2.1 Gaussian process regression

We give a brief introduction of GPs and regression for the context of the methods presented in this paper – for a fuller description please see Rasmussen & Williams (2006). A GP is a special case of stochastic process based on a multivariate Gaussian distribution (extended to an infinite number of variables), and is useful for both

regression and classification. In the context of regression, GPs allow us to infer probability distributions over non-parametric functions, affording highly flexible models, and are tractable since we are only concerned about evaluation at a finite number of outputs. We define the properties of the functions modelled by a GP via a mean function $m(\mathbf{x})$ and a covariance function $k(x_i, x_j)$, incorporating any prior knowledge of the form of the functions. We use the chosen kernel, $k(x_i, x_j)$, to compute covariance between any two inputs, thereby constructing the covariance matrix $\mathbf{K}$ of the GP. For an input vector $\mathbf{x}$, we assume our observations $\mathbf{y}$ are drawn from a multivariate Gaussian distribution:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{m}, \mathbf{K}). \tag{2}$$

It follows that the GP log-likelihood for N observations $\mathbf{y}$ is given by

$$\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}, \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{y}-\mathbf{m}) - \frac{1}{2}\log|\mathbf{K}| - \frac{N}{2}\log 2\pi, \tag{3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are the parameters of the mean and covariance functions, thus hyperparameters of the GP. By evaluation of this likelihood, we first sample the posterior distributions of the hyperparameters

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\theta})p(\boldsymbol{\phi}, \boldsymbol{\theta}), \tag{4}$$

where $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ is the prior on the hyperparameters. Having trained the GP, we are interested in prediction at test locations $\mathbf{x}_*$. Since the training and test sets are jointly Gaussian distributed, the predictive posterior distribution can be derived as

$$\mathrm{p}(\mathbf{y}_*|\mathbf{y}, k) = \mathcal{N}\left(\mathbf{K}_*^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{K}_*\right) \tag{5}$$

for $\mathbf{m} = \mathbf{0}$, where $\mathbf{K}_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$, the covariance function evaluated for pairs of test points and $\mathbf{K}_* = k(\mathbf{x}, \mathbf{x}_*)$ that for pairs of training and test points.

2.2 Application of GP regression to high-resolution spectroscopy

Although some underlying physics concerning possible absorbing species, temperature structure, and cloud formation is established, it is difficult to construct a suitable (physically motivated) model parametrization for high-resolution exoplanet spectra. By modelling with a GP, we acknowledge that there exists some correlation between flux uncertainties in wavelength, albeit locally, which can be specified via a parametrized covariance function. We assume little about the functional form of the spectra a priori, modelling each component spectrum of equation (1) as

$$f \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(\boldsymbol{\lambda}, \boldsymbol{\theta})), \tag{6}$$

where f is the spectral fluxes (e.g. F_{p}) and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)^{\mathrm{T}}$ is the distinct vector of wavelengths in the rest frame of the particular component. In theory, we could use a multidimensional GP to forward model the composite spectrum F_{obs}, as given in equation (1).

Though a GP is not a particularly natural choice of model for a spectrum, such an approach offers a number of anticipated benefits. First, whereas HRCCS only allows detection of the atmospheric chemical species included in the cross-correlation template, the proposed GP method affords an estimate of the functional form of the planetary spectrum, potentially facilitating detections of unsought species (and therefore unforeseen physics). In principle, this planetary spectrum could be directly fed into a retrieval framework in a similar way to low-resolution retrievals. That said, bypassing some of the available information, such as the location of planetary absorption features, is likely to cause a reduction in sensitivity compared to HRCCS. Further, using this methodology we are able to robustly propagate uncertainty estimates in a principled manner. This was a prevalent shortfall with standard HRCCS techniques, though we note that the cross-correlation to likelihood method now allows it, albeit in cross-correlation space (Brogi & Line 2019; Gibson et al. 2020). Analyses of high-resolution spectra also require continuum normalization for self-calibration. This results in the loss of broad-band variations and the planetary continuum, thus all measurements of absorption and emission features are relative. Though (e.g.) Brogi & Line (2019) and Line et al. (2021) show that there still remains enough information in the spectra to retrieve absolute abundances and temperature-pressure profiles, retaining the absolute planetary continuum would further help to constrain these atmospheric properties. Lastly, using GPs in this context allows modelling of correlations between data points; HRCCS methods assume the data are independent.

2.3 Sequential GP modelling

The number of operations required to compute the inverse and log-determinant of $\mathbf{K}$, needed for $\log \mathcal{L}$ in equation (3), scales as $\mathcal{O}(N^3)$. This is a limiting factor considering the typically large N for high-resolution observations. To mitigate this challenge, we use the PYTHON GP implementation CELERITE for our GP models, which makes use of specific, stationary forms of the covariance function for faster, $\mathcal{O}(N)$ evaluation (Foreman-Mackey et al. 2017). A stationary kernel is one that depends only on relative positions in input space rather than absolute positions. The form of the CELERITE kernel limits it to 1D inputs.[1] We therefore present a method to construct a model for and remove each component separately in this work, namely the Earth's transmission spectrum, $T(\boldsymbol{\lambda}_1, t)$, first and then the planetary absorption (or emission) spectrum, $F_{\mathrm{p}}(\boldsymbol{\lambda}_2)$ (see equation 1). Not only does this sequential method overcome the computational limitation, but also introduces modularity, allowing us to test each successive step in turn. We note that this method does not yet preserve the continuum, since we normalize the spectra prior to modelling. We assume that:

(i) the spectral components are each a realization of an independent GP;
(ii) the spectra are perfectly normalized and wavelength calibrated, and the instrument line spread function (LSF) is stable, so we are not concerned with instrumental effects;
(iii) the tellurics vary only with airmass;
(iv) the planet spectrum, F_{p} is constant in time;
(v) the stellar spectrum has been adequately removed prior to implementation of the GP framework.

In practice, these assumptions are unlikely to be strictly correct. We now proceed to examine each one in turn and discuss the likelihood and impact of it being violated.

(i) The assumption that each component can be treated as a GP has been addressed above already. Their independence is not in question, provided we are considering thermal emission. For reflected light, the planet's spectrum will depend on the star's spectrum, but this is not the case for the data sets considered in this work.

[1]Gordon, Agol & Foreman-Mackey (2020) extend CELERITE to 2D inputs, but only for very specific conditions that do not apply to the present problem.

(ii) If the wavelength calibration is significantly erroneous, this could lead to reduced or missed feature recovery. That said, this is a standard assumption in HRCCS. An unstable LSF would cause a variation in resolution of the observed planet spectrum with time, so should be assessed for each data set. The measured variation in resolving power for these data do not significantly affect the recovered GP spectrum.

(iii) While airmass is the dominant parameter controlling the time-dependence of telluric absorption, other factors are also important, such as variable precipitable water vapour. Furthermore, the dependence on airmass is no longer linear in the strong absorption regime. All in all, this is probably the most problematic assumption in our framework, and the reason why we also investigate other telluric correction methods in Section 3.3.

(iv) The planet's spectrum may change in time, particularly as the dayside rotates into view. First of all, the overall planet-to-star contrast should change, as the dayside of the planet rotates into or out of view. This would not be a problem when using the cross-correlation method to identify the planet's spectral features, but might need to be taken into account when forward-modelling the star and planet spectra simultaneously. Modelling them sequentially alleviates this problem, as we essentially lose the continuum information. Furthermore, as the day- and nightside can have very different temperatures, this can result in significant changes in atmospheric structure and composition, which would affect the features in the spectrum. We expect negligible variation in these data, considering the duration of observations and the relatively long rotational periods.

(v) This assumption was addressed in detail by Chiavassa & Brogi (2019). For the purposes of this work, while residual stellar features may impede our ability to recover the planetary spectrum to some extent, the effect of residual tellurics is much more significant.

Appropriate choice of kernel is important for any GP model; Rasmussen & Williams (2006) offer an extensive guide on model selection in ch. 5. For astrophysical applications, it is preferable to consider the physical processes producing the signal when choosing a suitable kernel. Atomic and molecular absorption is a somewhat predictable process but variation in abundance along the line of sight, weather conditions and presence of clouds and hazes can vary the occurrence, breadth and strength of absorption. In our case, the spectral features are altogether modelled via the covariance function (since $\mathbf{m} = \mathbf{0}$). The Matern class of covariance functions, $k_\nu(x_i, x_j)$ where ν defines the degree of differentiability of the output functions, produces somewhat rough behaviour compared to other widely used kernels such as the squared exponential kernel. Rajpaul, Aigrain & Buchhave (2020) assessed kernels for modelling of stellar spectra, finding the Matern-5/2 kernel to offer reasonable flexibility and smoothness. We choose to use the slightly rougher Matern-3/2 kernel because it (marginally) outperformed other kernels, available in the current CELERITE library, in a series of injection-recovery tests. It takes the form

$$k_{3/2}(\lambda_i, \lambda_j) = \sigma^2 \left(1 + \frac{\sqrt{3}}{\rho}|\lambda_i - \lambda_j|\right) \exp\left(-\frac{\sqrt{3}}{\rho}|\lambda_i - \lambda_j|\right), \quad (7)$$

where $\boldsymbol{\theta} = (\sigma, \rho)$ are hyperparameters corresponding to an amplitude and length scale. For this kernel, the covariance between data points separated by ρ in input space falls by 50 per cent. While we use the same kernel for both components (subsequently discussed) in this work, this is not obligatory for either the sequential method presented here or a hypothetical, future implementation where multiple components are modelled simultaneously.

2.3.1 *The telluric component*

High-resolution spectrographs are typically ground-based instruments, hence observations are subject to contamination by the Earth's transmission spectrum. The dominant absorbers are H_2O, OH, and O_2, though other species such as CH_4 and CO_2 contribute significantly in the NIR. It is vital to remove imprinted features from the spectra accurately to distinguish the true planet signals. The telluric spectrum is variable over a period of HRS observations due to weather conditions, variable water vapour abundance and airmass. Assuming a plane-parallel atmosphere, telluric transmission follows

$$T(\lambda, t) = \frac{I}{I_0} = \exp(-\tau(\lambda)a(t)), \quad (8)$$

where $I(\lambda)$ is the intensity, $I_0(\lambda)$ that above the atmosphere, $a(t)$ the airmass at time t and τ the optical depth at zenith ($a = 1$) (Noll et al. 2012). We therefore construct a telluric model

$$T(\lambda, t) = T_{\rm ref}(\lambda)^{a(t)}, \quad (9)$$

where $T_{\rm ref}(\lambda)$ is the Earth's transmission at zenith. A widely used technique in the field, which assumes the same treatment of tellurics, builds $T_{\rm ref}$ via linear regression to derive τ (Vidal-Madjar et al. 2010; Astudillo-Defru & Rojo 2013; Wyttenbach et al. 2015). Here, we model it with a GP. Given exponentiation is not an affine transform, we model the tellurics in log space. While we recognize that the propagated uncertainties (in log space) are no longer Gaussian, this does not have a significant impact on the result, likely due to the flexibility of GP models. We evaluate $\log\mathcal{L}$ (equation 3) on the time-average of observations $\mathbf{y} = \log(F_{\rm obs})/a(t)$, assuming all exposures share a common wavelength solution in the Earth's rest frame, and use the maximum likelihood estimate (MLE) to set the values of $\boldsymbol{\theta} = \{\sigma, \rho\}$. We treat each detector separately since the level of telluric absorption is expected to vary between them. For the newly defined $\mathbf{K}$, we compute the predictive distribution (equation 5), for $\mathbf{x}_*$ defined at each pixel on the detector. The predictive mean $\boldsymbol{\mu}_*$ is then used to construct $T_{\rm ref} = \exp(\boldsymbol{\mu}_*)$.

2.3.2 *Modelling the planet signal*

Once the spectra have been corrected for telluric absorption we attempt to find the planet signal within the noise of the residuals. We compute the equivalent wavelengths of each spectrum in the rest frame of the planet, having considered corrections for the Solar system barycentric velocity ($v_{\rm bary}$), velocity of the observed system ($v_{\rm sys}$), and the planet's radial velocity. For clarity, the resulting Doppler shift is given by the total planet velocity

$$v_{\rm p} = v_{\rm bary} + v_{\rm sys} + K_{\rm p}\sin(2\pi\varphi), \quad (10)$$

where φ is the planet orbital phases and $K_{\rm p}$ is the semi-amplitude of the planet radial velocity signature. Rather than analyse each spectrum individually, we combine the shifted spectra from all spectral orders to produce a composite 'spectrum' in the rest frame of the planet, by concatenating with the proper wavelengths in the planet rest frame. We model the planet spectrum with a second, independent GP. Though we are able to fit for $K_{\rm p}$, this method is not sensitive to the constant $v_{\rm sys}$ since we only work with relative wavelength shifts, with no knowledge of the position of features. We place flat priors on all three free parameters, $K_{\rm p} \sim \mathcal{U}(100, 200)\,{\rm km\,s^{-1}}$, $\sigma \sim \mathcal{U}(0, \infty)$, and $\rho \sim \mathcal{U}(d\lambda, \infty)$ where $d\lambda$ is the spectrograph resolution, and compute the GP log-likelihood (see equation 3). Exploration of the full joint posterior distribution [e.g. using a Markov chain Monte Carlo (MCMC) algorithm such as EMCEE (Foreman-Mackey et al.

Table 1. Summary of test data sets analysed in Section 3, where $t_{exposure}$ is the exposure time, $M_{spectra}$ is the number of separate exposures, and N_{SYSREM} gives the optimal number of SYSREM iterations found in Section 4.2 for each of the four detectors.

Object	Date	Phase coverage	$t_{exposure}$ (s)	$M_{spectra}$	N_{SYSREM}
HD 189733 b	2011-08-01	0.383–0.475	150	48	[4, 5, 4, 4]
51 Peg b	2010-10-16	0.36–0.42	42	166	[3, 3, 4, 2]
	2010-10-17	0.60–0.66	42	148	[2, 1, 5, 1]
	2010-10-25	0.49–0.54	42	138	[2, 5, 3, 3]

2013)] enables estimates of the three parameters with uncertainties. We use these estimates to construct the covariance matrix and finally condition the GP, obtaining the predictive mean planet spectrum. Since we do not expect significant correlation between test points located far apart in wavelength space, we condition the GP on the subset of data within 10ρ of each test position.

3 TESTS ON NIR ARCHIVAL DATA

3.1 Data format and prior reduction

To test our GP framework, we reanalyse archival observations of the widely studied HD 189733 b and 51 Peg b, taken on 2011 August 1 and over three nights, 2010 October 16th, 17th, and 25th, respectively. We attempt to recover the previously published, significant detections obtained by the standard methods introduced Section 1. Table 1 summarizes the details of the dayside observations; we refer the reader to the original analyses of these data in Birkby et al. (2013, hereafter BI13) and Brogi et al. (2013, hereafter BR13). In both cases, observations were taken with the CRyogenic Infra-Red Echelle Spectrograph (CRIRES; Kaeufl et al. 2004) on the Very Large Telescope, as part of the large ESO Program 186.C-0289. CRIRES imaged the spectra via four Aladdin II detectors, each 1024 × 512 pixels with gaps between each detector, with wavelength coverages 3.1805−3.2659 μm and 2.287−2.345 μm, and a resolution of $R = \lambda/\Delta\lambda \simeq 100\,000$. Extraction of the spectra and basic data reduction was completed via the CRIRES pipeline, v2.2.1 and v1.11.0 for the HD 189733 b and 51 Peg b observations, respectively.

In both cases the original authors grouped the spectra into matrices M spectra by N pixels, one for each of the four detectors on CRIRES. In this work, we begin the data processing and implementation of the GP framework with the pre-processed spectra, which were previously flat-fielded, bad pixel corrected, and background subtracted. We note that the spectra had also been aligned to a common wavelength grid in the Earth's rest frame; BI13 used the highest SNR spectrum as a reference with which to cross-correlate and BR13 compared the positions of the centres of telluric lines with the average spectrum.

3.2 Attempted recovery of the planet signal

3.2.1 HD 189733 b

HD 189733 is a K1V-type star with $T_{eff} = (4875 \pm 43)$ K, therefore we expect few strong stellar lines in the 3.2 μm domain. However, we assessed the telluric-corrected residuals at the locations of expected stellar lines with greater than 5 per cent absorption and found up to 2.5 times higher dispersion than the rest of the spectrum. In the original study, the authors did not explicitly remove any stellar features, as the PCA approach used to remove the tellurics was also able to capture the comparably weaker stellar lines, which shift by less than a pixel during the observations. Since our GP telluric removal assumes all stellar features are removed in order to scale by airmass, we remove them beforehand; we adopt a 1D stellar model from the PHOENIX stellar atmosphere model grid (Husser et al. 2013), interpolating[2] between the models for the literature values $T_{eff} = 4875$ K, $\log g = 4.56$ and metallicity [Fe/H] = 0.03 (listed in Table 2). We convolve the stellar model to the CRIRES resolution with a Gaussian kernel, scale according to the depth of two known stellar lines in the observations, and divide it out. Next, synonymous with typical high-resolution cross-correlation analyses, we correct for variation in throughput via continuum normalization; for each spectrum we fit a low-order polynomial to the top 10 per cent of fluxes in each detector, and divide it out. The normalization is not perfect here due to significant absorption in these spectra. Once normalized, we endeavour to remove the telluric contamination; water bands around 3.2 μm contribute to significant and even saturated telluric absorption, seen as vertical bands in the central, left-hand panel of Fig. 1. We divide out the predictive GP telluric model $T(\lambda, t)$ producing residuals as shown in the bottom, left-hand panel of Fig. 1. Subsequently, we flag any columns of the matrix in which the scatter of the data is systematically larger than the formal uncertainty from the CRIRES pipeline, and choose to inflate the propagated uncertainties accordingly. In some regions, the residuals showed a significantly larger scatter than the average, corresponding to regions of saturated tellurics. These regions are masked post telluric removal; we mask a column if its standard deviation is larger than a few factors of the median across the spectrum, with additional sigma clipping of individual points. We were left with an average of 52 per cent unmasked pixels across detectors 1, 3, and 4 but only 27 per cent for detector 2 given the severe atmospheric absorption between 3204.45 and 3222.45 nm. Having applied the masks, the average residual dispersion in continuum regions was 1.4 times the photon noise level.

We now attempt to recover the previously published planetary atmospheric detection, namely the H_2O planetary spectrum detected at 4.8σ in BI13. Before training and conditioning the GP for the planetary spectrum on the telluric-corrected residuals, we shift them into the planetary rest frame via equation (10) for an assumed $v_{sys} = -2.361\,km\,s^{-1}$, and v_{bary} calculated using BARYCORRPY (Wright & Eastman 2014; Kanodia & Wright 2018). We evaluate the GP predictive distribution over a test grid corresponding to one data point per resolution element of the spectrograph. In typical cross-correlation analyses, a detection is often quantified by dividing the values in the summed 2D CCF by the standard deviation across the entire matrix (or sometimes on a per spectrum basis), thereby providing a signal-to-noise detection significance. Since we do not utilize cross-correlation, we base our definition of a 'detection' on others factors: primarily we only consider a detection if the recovered

[2] All interpolations made using the SCIPY.INTERPOLATE package.

Table 2. Adopted parameters for both the HD 189733 and 51 Peg systems.

Parameter	Value	Reference
HD 189733		
$T_{\rm eff}$ (K)	4875 ± 43	Boyajian et al. (2015)
log(g)	4.56 ± 0.03	Boyajian et al. (2015)
Fe/H	-0.03 ± 0.04	Bouchy et al. (2005)
K_* (m s^{-1})	205 ± 6	Bouchy et al. (2005)
R_* ($R_\odot$)	$0.766^{+0.007}_{-0.013}$	Triaud et al. (2018)
$v_{\rm sys}$ (km s^{-1})	-2.361 ± 0.003	Bouchy et al. (2005)
HD 189733 b		
$R_{\rm p}$ (R_J)	$1.178^{+0.016}_{-0.023}$	Triaud et al. (2018)
$K_{\rm p}$ (km s^{-1})	154^{+14}_{-10}	BI13
51 Peg		
$T_{\rm eff}$ (K)	5793 ± 70	Fuhrmann, Pfeiffer & Bernkopf (1997)
K_* (m s^{-1})	55.65 ± 0.53	Wang & Ford (2002)
R_* ($R_\odot$)	$1.1609589^{+0.0222188}_{-0.0807567}$	Gaia Collaboration et al. (2018)
$v_{\rm sys}$ (km s^{-1})	-33.2 ± 1.5	BR13
51 Peg b		
$R_{\rm p}$ (R_J)	1.2	See Section 3.2.2
$K_{\rm p}$ (km s^{-1})	134.1 ± 1.8	BR13

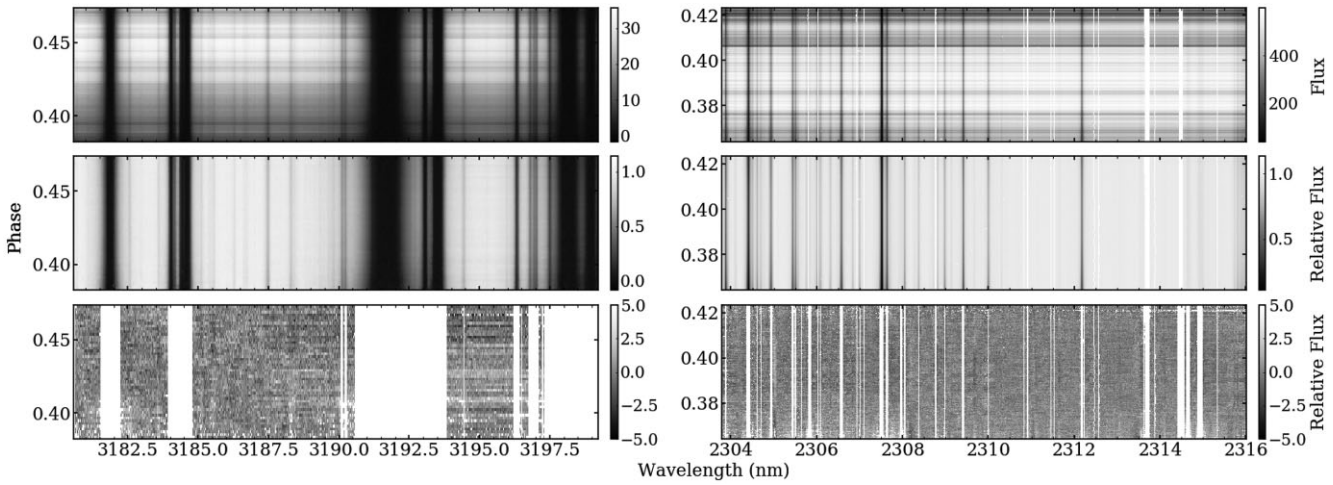

Figure 1. HD 189733 (*left*) and 51 Peg (*right*) spectra imaged on CRIRES detectors 1 and 2, respectively. *Top panel:* Reduced spectra matrix from BI13 and Chiavassa & Brogi (2019) post alignment, background subtraction, and bad pixel correction. *Middle panel:* Continuum-normalized spectra, showing vertical bands of telluric absorption. In the case of HD 189733, the stellar spectrum is removed between this and the next panel. *Bottom panel:* Residuals having removed the GP telluric model and masked appropriate columns, weighted by propagated uncertainties.

orbital velocity is in agreement with the published (or later injected) velocity. Applying this framework to the observed fluxes, the GP could not detect the true planet signal. We explore two potential reasons why the GP fails to recover the true signal: imperfect telluric removal, impact of which is discussed in Section 3.3, and the insensitivity of the GP to the noise levels in these data.

In applying GP regression we make no assumptions a priori with regard to the shape or form of the planetary spectrum, including the locations in wavelength of the absorption lines. The only assumptions we make are regarding the functional form of the covariance between data points, as given in equation (7). This potentially reduces the sensitivity of the GP method compared to HRCCS, wherein comparisons are made with a pre-defined template of absorption lines. To investigate the strength of signal which could be recovered, we inject model planet spectra at the previously detected radial velocity ($K_{\rm p}$, $v_{\rm sys}$), prior to the data reduction. Cross-correlating an H_2O + CO_2 model produced the highest detection significance in BI13, a planet spectrum model produced via line-by-line calculations with reference to the HITEMP line data base (Rothman et al. 2010; De Kok et al. 2013). We proceed with the H_2O-only model, as in BI13, since the inclusion of CO_2 only increased the detection significance by 0.3σ. This model assumes a baseline continuum temperature $T_1 = 500$ K with corresponding pressure $P_1 = 10^{-1.5}$, constant lapse rate and upper atmospheric temperature $T_2 = 1350$ K, and VMR(H_2O) $= 10^{-5}$. We scale the model, $F_{\rm model}$, according to the host stellar continuum and inject into the observed (pre-processed) spectra, $F_{\rm obs}$. We adapt slightly the approach of Brogi et al. (2013, 2014), and choose to scale the model according to the continuum of the observations, $F_{\rm cont}$, to avoid injecting additional telluric (and stellar) features. Our simulated observations are then

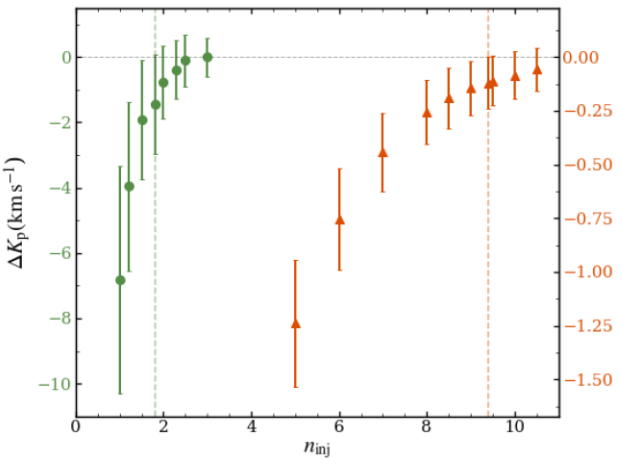

Figure 2. Difference between the injected and recovered K_p for HD 189733 b (green circles and left axis, injected $K_p = 154\,\mathrm{km\,s^{-1}}$) and 51 Peg b (orange triangles and right axis, injected $K_p = 134\,\mathrm{km\,s^{-1}}$). The coloured dashed lines indicate the smallest n_{inj} for which we obtain a detection ($n_{inj} = 1.8$ and 9.4).

given by

$$F_{sim} = F_{obs} + F_{cont}\left(n_{inj}\frac{F_{model}}{F_*}\left(\frac{R_p}{R_*}\right)^2\right), \quad (11)$$

where F_* is taken to be the blackbody estimate of the stellar continuum for the stellar effective temperature T_{eff}. Subsequently, these new spectra are analysed identically to the real spectra. We repeat the GP routine for different n_{inj}, monitoring the recovered K_p and uncertainty (the 1σ interval of the marginalized posterior distribution), as shown in Fig. 2. We recover a detection at $n_{inj} = 1.8$. The GP planet spectrum predictive mean and uncertainty for this injection strength is shown in the bottom, left panel of Fig. 3 for an unmasked portion of the spectrum, having adopted the MLE values $K_p = 152.55 \pm 1.53\,\mathrm{km\,s^{-1}}$, consistent with the injected $K_p = 154\,\mathrm{km\,s^{-1}}$, and the GP hyperparameters as shown in Fig. A1. Scaling the planet model according to equation (11) with $n_{inj} = 1$, the expected line strength (of the deepest lines) is 2.43×10^{-3}, which corresponds to about 20.1 per cent of the average standard deviation of the residuals. Note that BI13 used a scale factor $n_{inj} = -0.56$ to cancel out the real signal and recover a 0σ detection, thus computing an H_2O line contrast ratio of $(1.3 \pm 0.2) \times 10^{-3}$. This considered, $n_{inj} = 1.8$ corresponds to ∼3.2 times the detected cross-correlation signal strength.

3.2.2 51 Pegasi b

For the solar-type 51 Peg there are conspicuous, strong stellar CO lines in the vicinity of 2.3 μm; the absorption features in the composite spectrum are a convolution of telluric and stellar lines. These were challenging to remove, preventing detection of CO in the original analysis (BR13) using spectra from the third observing night (October 25th). For this night, the planet traversed superior conjunction, resulting in the CO lines of the planet and star overlapping. In the Chiavassa & Brogi (2019) reanalysis the authors used a more complete 3D stellar model, accounting for stellar convection, rather than the scaled solar 1D stellar model used previously. They then observed an increase in the detection significance of CO when including the spectra from October 25th. In this work, we analyse the stellar-corrected residuals of Chiavassa & Brogi (2019).

We normalize each spectrum by fitting a low-order polynomial to the top 10 per cent of pixels within each 10th of the spectrum to simultaneously capture any slight curvature across the order. At 2.3 μm, we expect less severe telluric absorption than at 3.2 μm, with features predominantly due to CH_4, which are known to be less variable than H_2O. Having divided out the continuum, and with stellar lines having been previously removed, only telluric absorption features common in wavelength and the Doppler shifting, relative planet atmospheric absorption remain. We remove the predictive GP telluric model, propagating the photon noise uncertainty estimates. The strong telluric line at $\lambda \simeq 2328.5$ nm, likely a feature of a third species, proved difficult to detrend so we choose to mask it. Additionally, post telluric removal we apply traditional channel masking using the same approach as detailed in Section 3.2.1.

Combining all three nights of data, once again we do not recover a detection of the true planetary signal. The template planet spectrum we inject is a CO + H_2O model based on a non-inverted temperature-pressure profile, with which cross-correlation produced a 5.9σ detection in BR13. This model was produced using the HITEMP data base (Rothman et al. 2010), assuming continuum temperature $T_1 = 1250$ K and pressure $P_1 = 0.1$ bar with constant lapse rate to top boundary pressure $P_2 = 1 \times 10^{-4}$ bar, $T_2 = 500$ K. The volume mixing ratios were $\mathrm{VMR}(H_2O) = 3 \times 10^{-4}$ and $\mathrm{VMR(CO)} = 1 \times 10^{-4}$; it should be noted that these absolute abundances were only weakly constrained since the models were not scaled for the cross-correlation in BR13. Though the authors also investigated the presence of CH_4, they did not observe a significant detection with pure CH_4 models. That said, at the time the line lists for CH_4 were incomplete and inaccurate for the purposes of high-resolution studies, and it would be important to re-evaluate conclusions regarding the presence of CH_4 using updated data bases (e.g. Hargreaves et al. 2020).

Scaling the model using equation (11), we obtain an expected absorption line strength of 1.48×10^{-4} relative to the continuum level, though again we note this may not be the true line strength of the signal in this data set. A contrast ratio of this scale is much smaller than the residual scatter of the GP telluric-corrected fluxes, corresponding to 3.33 per cent for detector 1. We scale the model spectrum using different values of n_{inj}, inject and process through the GP framework. For the radius of 51 Peg b, we adopt the average R_p of similar mass planets in the NASA exoplanet archive[3] ($R_p \sim 1.2\,R_J$), which agrees with the limits proposed by both Scandariato et al. (2021) and Birkby et al. (2017). For $n_{inj} = 9.4$ we recover $K_p = 133.88 \pm 0.12\,\mathrm{km\,s^{-1}}$; the right-hand panel of Fig. 3 shows the resulting GP from a subregion of detector 2 containing several planetary absorption lines, plotted over the telluric-corrected residual fluxes binned to one data point per resolution element. In this formalism R_p and R_* act as scaling factors in addition to n_{inj}, and use of a larger planetary radius would translate to a lower n_{inj}. Thus, to clarify, the injected model has an average line strength of $\sim 1.39 \times 10^{-3}$, a similar strength to the real HD 189733 b signal. We note that the real signal exists in the data, with a cross-correlation detection uncertainty of $\sim 2\,\mathrm{km\,s^{-1}}$. Therefore, though we know precisely the velocity at which we inject the model, it may be that the real signal is offset. At low n_{inj} then, it is plausible that the injected and (slightly offset) real signals blend together, distorting the shape of

[3] https://exoplanetarchive.ipac.caltech.edu/

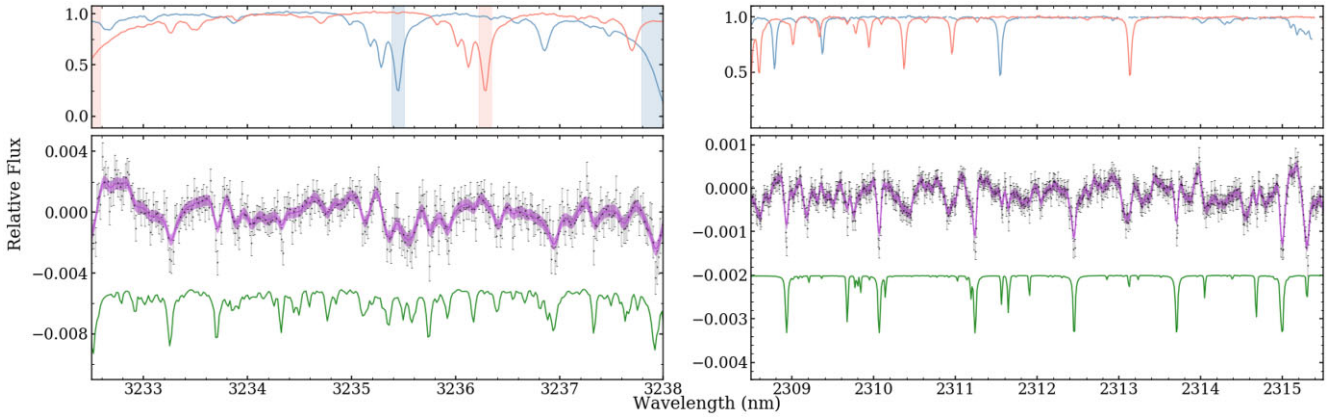

Figure 3. Smallest model injection GP recoveries. *Top panels:* Normalized average spectra plotted in the planet rest frame, dominated by telluric absorption features, with those corresponding to minimum and maximum v_p (see equation 10) shown in blue and red, respectively. Any masked wavelength channels are indicated by shaded regions. *Bottom panels:* Sub-section of the GP telluric-corrected residual fluxes, binned to CRIRES resolution (again in the planet rest frame) in black, with GP predictive mean and posterior uncertainty (1σ interval) overplotted in purple. The injected model spectrum is plotted in green, vertically offset for reference. *Left-hand panels:* Recovered HD 189733 b H_2O spectrum with injection strength $n_{inj} = 1.8$. *Right-hand panels:* Result for 51 Peg b, having injected a CO + H_2O model with strength $n_{inj} = 9.4$.

the recovered signal as well as the recovered K_p, therefore impeding a detection.

3.2.3 Radial velocity semi-amplitude estimation

As explained in Section 2.3.2, we obtain estimates of K_p and 2 hyperparameters with this method; Fig. A1 shows that the K_p estimates are tightly constrained. Inherently, the GP for the planet component assumes that the data have been perfectly detrended, so what remains is only white noise and the planet signal. This considered, it follows that the theoretical constraining limit of K_p is a combination of the spectrograph resolution, SNR, range of phase coverage, and number of (detectable) lines. In reality, however, residual stellar and telluric features exist (with negligible intra-night radial velocity) and become dominant for low n_{inj}, wherein the recovered K_p is seen to diverge from the injection velocity (Fig. 2). Though it should be noted that the formal uncertainties from HRCCS are likely not the smallest achievable, since the intermediate interpolation steps have an associated noise contribution, our uncertainty estimates for K_p are smaller than expected. To proceed to reliable hypothesis testing, en route to enabling standalone GP detections, we would require a more realistic noise model.

We also see a difference in constraining power of K_p between the two data sets. From Table 2, we see that the K_p estimates from HRCCS analyses have maximal uncertainties of 14 km s^{-1} and 1.8 km s^{-1}, with those for HD 189733 b a factor of $\sim$7.8 $\times$ larger. There are a few conceivable causes for this difference: dynamics and other effects may cause broadening of the spectral lines and thus the radial velocity signature, but the higher signal-to-noise of the 51 Peg b data due to wider phase coverage is likely to be a dominant factor. Since these are intrinsic to the data, we therefore expect a similar factor in our analyses. There may have also been slight differences in the methods used to estimate K_p from the cross-correlation maps by the authors of BI13 and BR13. Additionally, the CO lines of 51 Peg b's spectrum are rather more distinct and well defined than the spectrum of HD 189733 b, which is likely to significantly aid the detection.

3.3 Using other detrending algorithms

As mentioned in Section 3.2.1, poor telluric modelling using GP regression may contribute to inability to detect the buried planet signals. In order to keep the GP model sufficiently simple for the sequential method implemented here, we have only allowed temporal telluric variation corresponding to airmass. In reality telluric behaviour is more complex than this, hence notoriously difficult to model accurately in the NIR, particularly in the case of saturated and non-water-based absorption lines. An additional cause may be the interference of lines from the different components in equation (1). In the case where stellar, telluric and/or planet lines overlap, successive removal will partially affect the other components in the relevant exposures; this issue could potentially be improved with the hypothesized simultaneous GP approach mentioned in Section 2.2.

To test the contribution of imperfect telluric removal to insensitivity of the GP method to the real signal, we employ the original detrending methods, SYSREM, and an airmass detrending approach, which enabled detections via the cross-correlation method. We replicate the detrending methods of BI13 and BR13, with some slight alterations as detailed in the following sections. Doing so allows us to directly compare the sensitivity of the GP to the cross-correlation approach used in those publications for identifying the planet signal, by eliminating the impact of the GP telluric removal.

3.3.1 SYSREM

Compared to the rapidly Doppler shifting planetary signal (e.g. 103–24 km s^{-1} for the HD 189733 b observations) the telluric and stellar spectra appear quasi-stationary, experiencing only sub-pixel shifts over the course of an observing night. Analogous to other SYSREM implementations, we divide through the mean flux in each pixel and subtract unity, leaving only residual temporal variation (BI13). We note that it is necessary to apply masks to regions of saturated telluric absorption prior to input; for these tests we mask the same columns of the spectral matrix as BI13. We run the SYSREM algorithm for the same number of iterations, N_{SYSREM}, for detector 1 and 3 as used in BI13 ($N_{SYSREM} = 8$ and 1 iteration, respectively). We also

include detectors 2 and 4, unused in BI13, employing $N_{\rm SYSREM} = 5$ for each. We observe a non-detection for the SYSREM-treated fluxes. Separately, we test $N_{\rm SYSREM}$ found using our own optimization process (explained further in Section 4.2), but still do not achieve a detection. We inject the H_2O model spectrum in the same way as Section 3.2.1 and again investigate the lowest $n_{\rm inj}$ for which we observe a detection. At $n_{\rm inj} = 1.4$, we recover an agreeable planet orbital velocity, $K_{\rm p} = 155.62 \pm 1.76\,{\rm km\,s^{-1}}$. SYSREM visibly better suppresses the noise, thus it is unsurprising that smaller signals are detectable.

3.3.2 Airmass-detrending approach

We employ the linear regression technique presented and used by Brogi et al. (2012, 2014) and Schwarz et al. (2015) to remove the telluric lines imprinted on the 51 Peg spectra. After normalizing the spectra, we identify and remove any airmass-related variation in each pixel before sampling the residuals at user-defined λ_s, known strong telluric line positions, for any second-order effects, and applying a high-pass filter as discussed in Section 1. These steps replicate the approach outlined in BR13, with the exception of normalizing by column variances, a final step commonly used in cross-correlation analyses to 'down-weigh' particularly noisy pixels. We choose to omit this step since it alters the variance of the data. Instead, we mask noisy channels and apply sigma clipping during the final high-pass filtering using the masking algorithm described in Section 3.2.1. Fig. 4 shows the spectral matrix residuals from a single night (October 16th) after each step of the reduction, with the third panel showing the residuals having first removed the airmass trend. Again, we do not achieve a detection of the planet signal having treated the fluxes with this sequence of detrending algorithms. We varied λ_s from those used in BR13 but to no avail. Having injected the CO + H_2O model spectrum with $n_{\rm inj} = 8.5$, we achieve a marginal posterior mean $K_{\rm p} = 134.13 \pm 0.14\,{\rm km\,s^{-1}}$.

The ability of the GP to recover a marginally weaker planet signal both here and using SYSREM suggests that the GP telluric model is underperforming, though we note that here the aforementioned $n_{\rm inj} = 8.5$ corresponds to an average ${\rm SNR_{line}} \sim 0.36$ (of the strongest lines) when compared to the noise of the detrended residuals, higher than ${\rm SNR_{line}} \sim 0.29$ for the GP-treated fluxes. Significant residuals are visible in the channels containing stronger telluric lines after the first pass of this linear regression detrending technique (third panel of Fig. 4), illustrating that the behaviour of tellurics is more complex than assumed. This further supports that the airmass scaling assumption is likely to be too simplistic and the predominant shortfall of the GP sequential method for detecting small signals.

4 QUANTITATIVE ANALYSIS OF THE IMPACT OF DETRENDING

It has been shown that detrending methods used in past HRCCS analyses can alter the intrinsic planet signal. So, in order to avoid biasing the exoplanet atmospheric retrieval, it is important to replicate any alterations by subjecting the cross-correlation template to the same detrending process (Brogi & Line 2019; Serindag et al. 2021). Still, it is difficult to identify the exact effect of detrending as the planet spectrum is unseen. Alternatively, masks can be applied prior to detrending, to protect regions expected to contain planet absorption or emission features (Schwarz et al. 2015). Not only does this require input knowledge of atmospheric content, but it is also non-ideal to rid parts of the domain which may contain valuable information. Our

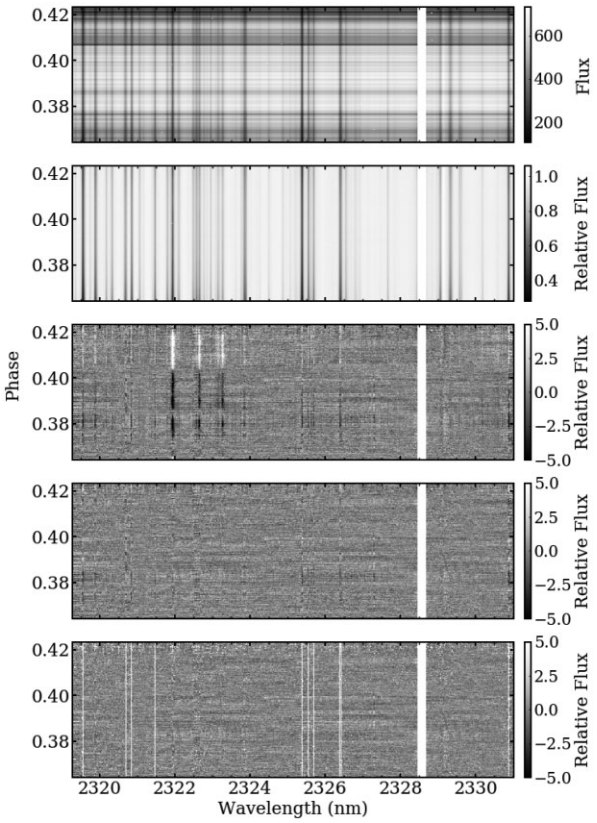

Figure 4. Data reduction of 51 Peg spectra from October 16th imaged on CRIRES detector 3. The strong telluric line at $\lambda \simeq 2328.5$ nm is masked throughout. *Top panel:* Stellar-subtracted fluxes from Chiavassa & Brogi (2019). *Second panel:* Continuum-normalized fluxes. The latter three panels show the results of successive routines of the linear regression telluric removal technique outlined in Section 3.3.2. Noisy channels are masked post detrending in the final panel.

methodology allows us to observe and measure the modifications to the planet spectrum itself.

4.1 Defining detection metrics

We develop four metrics to assess the impact of the telluric removal on the planet signal, having considered which properties of a spectrum are important to preserve for the purposes of retrievals. When characterizing a planetary atmosphere, we need to distinguish features in absorption from those in emission, indicative of the presence of temperature inversions in the upper atmosphere. Since here we are working with injected spectra which are known to contain only absorption features, we reserve this assessment for future work. When comparing the recovered 'planetary spectrum' to the injected model spectrum, we first assess if absorption features are observed at the same wavelength locations. Then, we measure the depths of absorption lines relative to the 'pseudo-continuum', which remains after detrending. These are important for deciphering atomic and molecular abundances, and are also affected by sources of broadening such as rotation. We identify any features greater than 2σ away from the continuum in the injected model spectrum, and fit the located absorption lines with a Gaussian profile to measure their depths, widths, and positions. We note that dynamics in the planetary atmosphere can distort spectral features, resulting in asymmetric profiles. For the test cases in this work, we observe

little distortion of the line profiles in the injected models, therefore while we measure the full width at half-maximum (FWHM) of the lines, we leave assessment of line shape to future work and assume symmetric features. In addition, we are interested in reproducing the planet 'pseudo-continuum', thus minimizing the presence of spurious features in the telluric-corrected fluxes. Therefore, as a third metric we select regions of continuum and assess dispersion. We compute a final combined metric which considers that each metric should be as close to unity as possible, weighting by uncertainty where appropriate and normalizing before combining. To summarize, we measure the following four properties:

(i) line depth;
(ii) line FWHM;
(iii) dispersion in continuum regions;
(iv) a metric that combines the above.

For the model comparison, we inject the template spectrum into a white noise matrix with a dispersion set to the photon noise level, at the published radial velocity of the planet. Then, the model comparison spectral properties are those measured having shifted the 'noise + injected model' matrix to the planet rest frame and binned to the resolution of CRIRES.

4.2 Tuning detrending parameters

Though SYSREM is a powerful detrending tool, it can be easy to degrade the planetary signal. In most cases, it is likely that planet spectral features experience sub-pixel shifts between exposures – SYSREM can begin to remove these sub-pixel common modes once the telluric spectrum has been removed for higher (user-specified) number of iterations. There is therefore a fine balance between removing as much of the stellar and telluric contamination as possible while retaining the planet signal. The common method used in the literature to determine the optimal number of SYSREM iterations, $N_{\rm SYSREM}$, is to inject a model planet spectrum at the expected radial velocity ($K_{\rm p}$, $v_{\rm sys}$), and select $N_{\rm SYSREM}$ that corresponds to the maximum recovery significance, typically the SNR of the peak in a cross-correlation map (Birkby et al. 2017; Nugroho et al. 2017; Kesseli et al. 2020). The noise is estimated as the standard deviation of CCF values in a user-defined region, located away from the predicted peak. There have been some concerns that this SNR is not a robust metric for the purpose of tuning detrending parameters; Cabot et al. (2019) showed use of SNR as an optimization metric to be susceptible to false-positive detections due to the presence of spurious features in the cross-correlation maps. This has also created caution over optimization of iteration number on an order-by-order basis, leading some to opt for uniform SYSREM use across spectral orders (Nugroho et al. 2017, 2020).

To determine the optimal number of SYSREM iterations for recovering the planet signal, we first inject a signal larger than the strength detected in Section 3.3.1. We train the GP only on the residuals following three iterations applied to each detector. We then run SYSREM for eight iterations, saving the residuals after removal of each principal component and evaluate the GP, having used the globally optimized covariance function and $K_{\rm p}$, for each residual-flux matrix. We make use of the three metrics developed in Section 4.1 to assess line recovery of the deepest planet absorption lines. The amplitude and FWHM recovery-injection ratios are averaged over all the selected absorption lines. In the case of the 51 Peg b data set, we repeat this process for each night separately; it is expected that telluric contamination will not only vary between detectors but also significantly between nights of observations. The measured

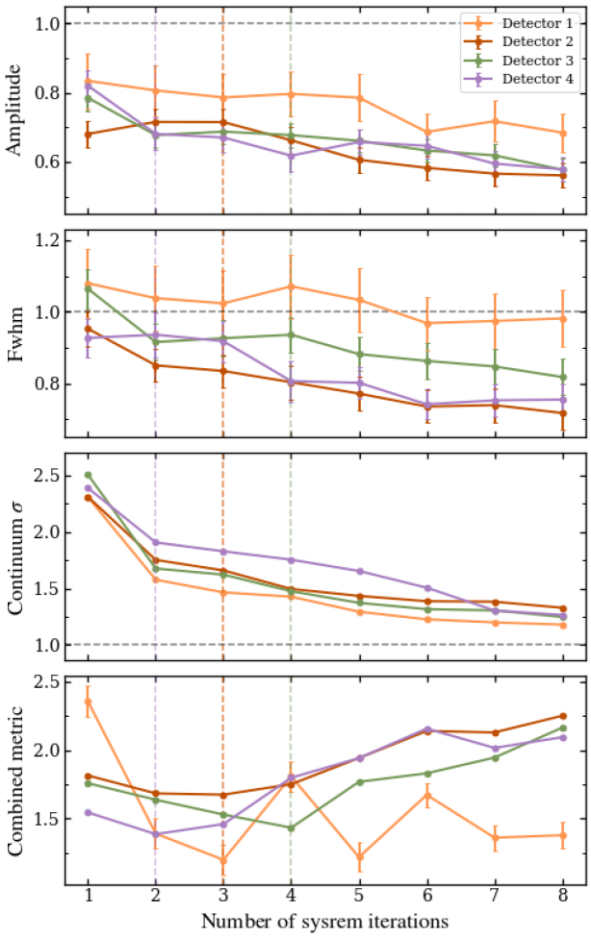

Figure 5. Assessment of planet ($CO + H_2O$) spectrum recovery for 51 Peg b observations taken on October 16th using different number of SYSREM iterations, having injected a signal of strength $n_{\rm inj} = 14.0$. Each of the three metrics, (i)–(iii) in Section 4.1, is shown individually in the first three rows. *Bottom panel:* Minimization of combined metric having considered line amplitude, FWHM, and continuum standard deviation recovery of the planet spectrum, as compared to the injected (noisy) model. Optimization is performed on each detector independently as indicated by the different colours.

properties for the first night of 51 Peg b observations (October 16th) are shown in Fig. 5. We select the $N_{\rm SYSREM}$ that minimizes the combined metric (bottom panel of Fig. 5). In the case of slight ambiguity concerning the minimum difference between injection and recovery, we select the lower $N_{\rm SYSREM}$. The optimum values used going forward are summarized in Table 1. Since the method presented here does not use the SNR of cross-correlation maps as a detection metric, it bypasses the issue of noise optimization, thus reducing the risk of biasing the detection by optimization of SYSREM for each detector separately.

4.3 Comparison of tellurics correction methods

To compare the three methods of telluric removal, we process the same set of fluxes, with a model injection of the same strength, and ensure the masking is identical. Cabot et al. (2019) showed the impact of variation in mask level on detection significance – we do not investigate this explicitly in this section. We run the SYSREM algorithm for an optimal number of iterations as indicated

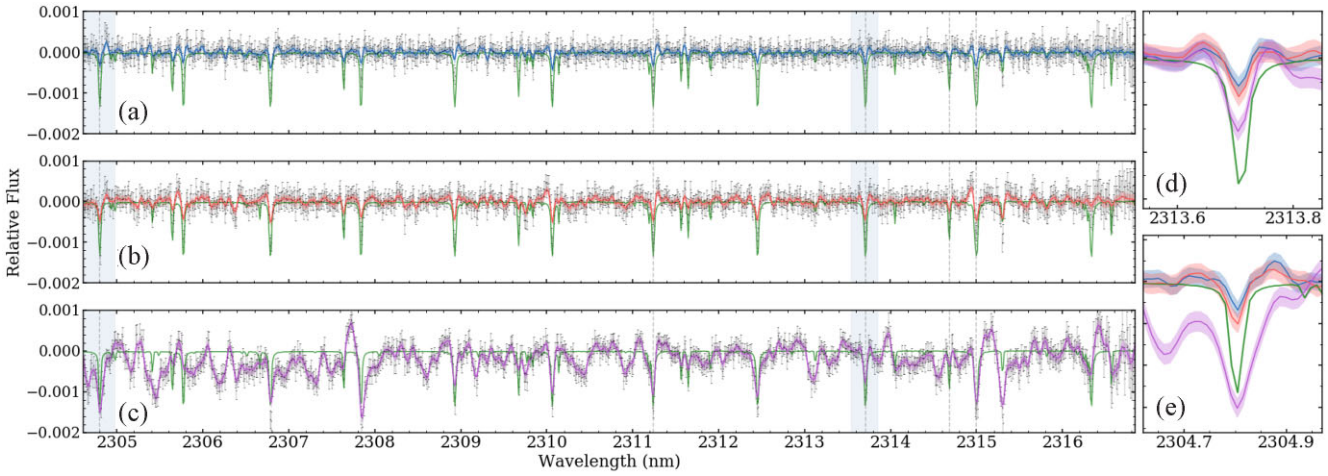

Figure 6. Binned linear regression-corrected 51 Peg residuals (panel a), SYSREM-corrected fluxes (panel b) and having removed a GP telluric model (panel c), in the planetary rest frame, with injected CO model of strength $n_{\rm inj} = 9.5$ (shown in green for reference). The GP regressed on the (full set of) residuals is overplotted in each case. The full resolution residuals have a large dispersion, thus to maintain clarity they are not shown here. The dashed line indicates the locations of the lines given in Table 3. Please note that panels (a–c) show three separate rather than sequential processes. Panels (d) and (e) are subsections of the wavelength domain around $\lambda \simeq 2313.709$ and $\lambda \simeq 2304.804$ nm, respectively, showing the GP predictive mean regressed on the 3 sets of residual fluxes. The corresponding regions are shaded in blue in panels (a–c).

in Table 1 removing each systematic component successively; the optimization process is explained in Section 4.2. We treat spectra from each detector (and each night) independently. Some authors opt to divide the fluxes through by a global model, the summation of all the identified principal components (Gibson et al. 2020; Merritt et al. 2020). We do not observe a significant difference in the recovered planet spectrum using this method instead, at these injection strengths.

Both SYSREM and the airmass fitting approach outperform the GP telluric model in terms of removal of stationary contaminating features, showing lower dispersion in the telluric-corrected residuals, and consequently increasing the ratio of expected line contrast to noise. The HD 189733 observations suffered saturated telluric absorption for a large proportion of the wavelength domain, up to 70 per cent for detector 2. All three methods struggled to model these and we had to apply significant masking. Despite the reduced spectral coverage, there were a few regions containing several absorption features (covering ranges of 1–2 μm) that enable comparison of line recovery. As seen in Section 3.3.1, SYSREM enabled recovery of a smaller (injected) planetary signal than the GP sequential method, and similarly using the linear regression method with the 51 Peg spectra (Section 3.3.2). That said, we observe a clear reduction in amplitude of the line features in the corrected fluxes when compared to the original injected model amplitudes. To quantitatively compare the impact of the three telluric detrending methods used, we measure the aforementioned line properties after applying the telluric correction, shifting the corrected fluxes to the planet rest frame by the MLE $K_{\rm p}$ (obtained having implemented our GP modelling of the planet spectrum), and binning the residuals to one data point per spectrograph pixel. We show an example result applying each method to the 51 Peg spectra in Fig. 6 for $n_{\rm inj} = 9.5$, which includes the regressed GP overplotted for interest. For this example, the recovered line metrics for some notable lines and values averaged across the 10 deepest lines are given in Table 3, along with the measured continuum standard deviation of a set of pre-defined spectral regions.

Both SYSREM and the airmass linear regression methods perform poorly with regard to maintaining the strength of absorption lines. With the overall variance as a metric, sub-pixel shifts will be identified and removed with these methods; thus, the wings and depths of lines are degraded. Meanwhile, generally we find that the GP telluric model enables better retention of the line depths, and the GP telluric correction reproduces a noisy continuum, greater than three times the standard deviation of the 'noise + injected model' for the example given in panel (c) of Fig. 6 ($n_{\rm inj} = 9.5$). In addition to averaging these metrics across the spectrum, we assess the wavelength dependence. We identify planet absorption lines that satisfy differing conditions: those that consistently lie in regions with high atmospheric transparency (>90 per cent) and those that Doppler shift across strong telluric lines (<80 per cent), traversing the same spectral region in 10–90 per cent of exposures.

Measuring the amplitudes and FWHM ratios for the lines in each subgroup, we observe that SYSREM and the airmass linear regression residuals cultivate wavelength-independent line recovery, producing consistent ratios in regions of continuum and regions of strong telluric absorption. This is not the case with the recovered lines following the GP telluric correction. Continuing with the example of Fig. 6 and Table 3, of the 10 absorption lines measured, four lines are located in regions with more than 95 per cent telluric transmission: $\lambda \simeq 2311.241$, 2313.709, 2314.683, and 2314.999 nm. For reference, a portion of the average telluric spectrum for this detector can be seen in the top right panel of Fig. 3, plotted in the planetary rest frame to show relative position for the full duration of observations. Shown in the bottom right panel, the line located at $\lambda \simeq 2304.804$ nm traverses a deep telluric line over more than 60 per cent of the total exposures. The recovered to injected ratios for each of these individual lines are included in Table 3. While the example of detector 2 for the 51 Peg spectra is highlighted here, it is representative of results observed from other detectors. In general, we find that planet spectral lines that consistently lie in regions of light telluric absorption are well recovered, whereas the recovered line profiles are distorted if deep telluric lines are nearby.

We conclude that planet lines that are not in close proximity to telluric features in a large number of exposures are more reliably and accurately recovered with the sequential GP method, compared to

Table 3. Line recovery having implemented different methods of telluric correction. All metrics given as the ratio between those measured in the binned telluric-corrected fluxes and in the injected CO model. Example given for 51 Peg b, detector 2, $n_{\rm inj} = 9.5$ (all observing nights).

Line λ (nm)	Airmass linear regression	SYSREM	GP telluric correction
Amplitude			
Average	0.629 ± 0.038	0.680 ± 0.038	1.138 ± 0.053
2304.804	0.833 ± 0.180	0.919 ± 0.139	1.575 ± 0.227
2311.241	0.655 ± 0.154	0.692 ± 0.118	1.073 ± 0.194
2313.709	0.576 ± 0.121	0.628 ± 0.095	0.923 ± 0.125
2314.683	0.652 ± 0.164	0.636 ± 0.184	0.716 ± 0.247
2314.999	0.534 ± 0.108	0.549 ± 0.103	1.157 ± 0.146
FWHM			
Average	0.739 ± 0.049	0.808 ± 0.047	1.123 ± 0.065
2304.804	0.751 ± 0.152	0.805 ± 0.131	1.185 ± 0.239
2311.241	0.724 ± 0.191	0.758 ± 0.147	1.219 ± 0.264
2313.709	0.603 ± 0.133	0.637 ± 0.102	0.741 ± 0.120
2314.683	0.941 ± 0.241	0.923 ± 0.269	0.801 ± 0.355
2314.999	0.730 ± 0.160	0.795 ± 0.147	1.264 ± 0.174
Continuum standard deviation			
	1.622	1.726	3.747

the other two detrending methods used. One plausible use of these findings would be to incorporate some GP modelling in an HRS data reduction pipeline, for the purpose of informing the detrending rather than implementing it. For example, one could take the average line depth measured using the GP framework, having excluded lines adjacent to telluric lines, and use it as a scaling factor for the recovered planetary spectrum. This would then combine the strengths of both approaches to extract the most accurate planetary spectrum possible. This is only a proposal here; we stress that our aim in this work is not to put forward a competitive, stand-alone telluric detrending method, but to test a sequential GP framework with the aim of future extension to a simultaneous GP modelling method.

5 CONCLUSIONS

In this work, we have proposed a novel technique for analysis of high-resolution spectra, using GP regression to sequentially model the telluric and planet spectra directly. The Bayesian approach affords an estimate of planet orbital velocity in addition to the planet spectrum estimate, both with robust uncertainties. We made a number of simplifying assumptions for our models in order to attain efficient computation, notably only allowing airmass variation of the telluric model. Nevertheless, the simplistic, sequential framework has enabled useful tests en route to construction of a comprehensive GP treatment. Though we produced no detection on real data sets, we successfully recovered planet absorption line positions and shapes for injected signals slightly larger than the nominal detected signals from cross-correlation analyses. The method here was developed specifically for dayside exoplanet spectra; however, a similar procedure could be used for transit observations, with alterations considering the Rossiter–Mclaughlin and other stellar effects.

It is clear that the efficacy of the GP routine to retrieve the planet signal is dependent on the performance of the telluric removal for the particular data sets in this work. We investigated the impact of different detrending methods on injected planetary signals, assessing recovered absorption line profiles. Standard techniques, SYSREM and the aforementioned linear regression method, consistently degrade the sought planetary signal, both truncating the absorption features and the wings of lines, which in the most severe case could prevent detection of weaker lines. Any high-resolution analyses should apply these techniques with caution and take care to mimic the effect of detrending when comparing to models. Though SYSREM is very effective at removing the telluric imprint, it has the potential to be quite aggressive, thus requiring careful tuning. It is then difficult to choose the optimum number of iterations to remove as much of the telluric contamination as necessary, to reveal the 'true' planet signal while completely retaining that signal. Additionally, we observed that all absorption lines were uniformly affected across the spectrum with both standard detrending methods. In contrast, the line recovery of the combined sequential GP methodology was wavelength dependent, fairing better in regions of greater telluric transparency compared to regions in close proximity to strong telluric lines. The downfall is likely to be the simplistic airmass scaling of the telluric model.

Use of GP regression in this context offers a number of advantages; we obtained detections without the requirement of any prior (exact) knowledge concerning the shape or form of the spectra for construction of a parametrized model. Acquiring an estimate of the planet spectrum is greatly beneficial, and may further help with high-resolution atmospheric retrievals. We note that our current method requires continuum normalization in the data reduction; thus, we do not retain the absolute planetary continuum. This could potentially be rectified via extension to a simultaneous GP fitting. An additional benefit of the method we presented is the natural acquisition of a (GP) likelihood function, thus disposing the need for a conversion tool to output a statistically meaningful detection. Though for the signal-to-noise ratio of these data sets our current method does not have sufficient sensitivity to recover the real planet signals, the prospects of this technique are promising. With the proposed advancements, GPs may offer a data-driven, potential route forward in analysis of high-resolution spectra.

ACKNOWLEDGEMENTS

We thank the anonymous referee for their comments towards improving the clarity of this manuscript. AM acknowledges support from the UK Science and Technology Facilities Council (STFC). SA and JLB acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation

programme under grant agreement numbers 865624 and 805445. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. Based on observations collected at the European Southern Observatory under ESO programme 186.C-0289.

DATA AVAILABILITY

The data underlying this article are publicly available through the ESO Science Archive Facility.

REFERENCES

Aigrain S., Parviainen H., Pope B. J. S., 2016, MNRAS, 459, 2408
Angus R., Morton T., Aigrain S., Foreman-Mackey D., Rajpaul V., 2018, MNRAS, 474, 2094
Astudillo-Defru N., Rojo P., 2013, A&A, 557, A56
Beltz H., Rauscher E., Brogi M., Kempton E. M.-R., 2021, AJ, 161, 1
Birkby J., 2018, Exoplanet Atmospheres at High Spectral Resolution, preprint (arXiv:1806.04617)
Birkby J., De Kok R. J., Brogi M., de Mooij E. J., Schwarz H., Albrecht S., Snellen I. A., 2013, MNRAS, 436, 1980 (BI13)
Birkby J., De Kok R. J., Brogi M., Schwarz H., Snellen I. A., 2017, AJ, 153, 138
Bouchy F. et al., 2005, A&A, 444, L15
Bourrier V. et al., 2020, A&A, 635, A205
Boyajian T. et al., 2015, MNRAS, 447, 846
Brogi M., Line M., 2019, AJ, 157, 114
Brogi M., Snellen I. A., De Kok R. J., Albrecht S., Birkby J., De Mooij E. J. W., 2012, Nature, 486, 502
Brogi M., Snellen I. A., De Kok R. J., Albrecht S., Birkby J., De Mooij E. J. W., 2013, ApJ, 767, 27, (BR13)
Brogi M., De Kok R. J., Birkby J., Schwarz H., Snellen I. A., 2014, A&A, 565, A124
Cabot S. H. C., Madhusudhan N., Hawker G. A., Gandhi S., 2019, MNRAS, 482, 4422
Casasayas-Barris N. et al., 2019, A&A, 628, A9
Chiavassa A., Brogi M., 2019, A&A, 631, A100
Czekala I., Mandel K. S., Andrews S. M., Dittmann J. A., Ghosh S. K., Montet B. T., Newton E. R., 2017, ApJ, 840, 49
De Kok R. J., Brogi M., Snellen I. A., Birkby J., Albrecht S., De Mooij E. J. W., 2013, A&A, 544, 82
Ehrenreich D. et al., 2020, Nature, 580, 597
Flowers E., Brogi M., Rauscher E., Kempton E. M.-R., Chiavassa A., 2019, AJ, 157, 209
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, ApJ, 154, 220
Fuhrmann K., Pfeiffer M. J., Bernkopf J., 1997, A&A, 326, 1081
Gaia Collaboration, 2018, A&A, 616
Giacobbe P. et al., 2021, Nature, 592, 205
Gibson N. P., Aigrain S., Roberts S., Evans T. M., Osborne M., Pont F., 2012, MNRAS, 419, 2683
Gibson N. P. et al., 2020, MNRAS, 493, 2215
Gordon T. A., Agol E., Foreman-Mackey D., 2020, AJ, 160, 240
Gullikson K., Dodson-Robinson S., Kraus A., 2014, AJ, 148, 53
Hargreaves R. J., Gordon I. E., Rey M., Nikitin A. V., Tyuterev V. G., Kochanov R. V., Rothman L. S., 2020, ApJS, 247, 55
Hoeijmakers H. J. et al., 2020, A&A, 641, A123
Husser T.-O., Wende -Von Berg S., Dreizler S., Homeier D., Reiners A., Barman T., Hauschildt P. H., 2013, A&A, 553, A6
Kaeufl H.-U. et al., 2004, in Moorwood A. F. M., Masanori I., eds, Proc. SPIE Conf. Ser. Vol. 5492, Ground-Based Instrumentation for Astronomy. SPIE, Bellingham, p. 1218
Kanodia S., Wright J., 2018, Res. Notes AAS, 2, 4
Kesseli A., Snellen I. A., Alonso-Floriano F. J., Molliere P., Serindag D. B., 2020, AJ, 160, 228
Langeveld A. B., Madhusudhan N., Cabot S. H. C., Hodgkin S. T., 2021, MNRAS, 502, 4392
Line M. R. et al., 2021, Nature, 598, 580
Madhusudhan N., Seager S., 2009, ApJ, 707, 24
Mazeh T., Tamuz O., Zucker S., 2007, in Afonso C., Weldrake D., Henning Th., eds, ASP Conf. Ser. Vol. 366, Transiting Extrasolar Planets Workshop. Astron. Soc. Pac., San Francisco, p. 119
Merritt S. R. et al., 2020, A&A, 636, A117
Noll S., Kausch W., Barden M., Jones A. M., Szyszka C., Kimeswenger S., Vinther J., 2012, A&A, 543, A92
Nugroho S. K., Kawahara H., Masuda K., Hirano T., Kotani T., Tajitsu A., 2017, AJ, 154, 221
Nugroho S. K., Gibson N. P., De Mooij E. J. W., Herman M. K., Watson C. A., Kawahara H., Merritt S., 2020, ApJ, 898, 31
Pino L. et al., 2020, ApJ, 894, L27
Piskorz D. et al., 2016, ApJ, 832, 131
Piskorz D. et al., 2017, AJ, 154, 78
Piskorz D. et al., 2018, AJ, 156, 133
Rajpaul V., Aigrain S., Osborne M. A., Reece S., Roberts S., 2015, MNRAS, 452, 2269
Rajpaul V. M., Aigrain S., Buchhave L. A., 2020, MNRAS, 492, 3960
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA
Rodler F., Lopez-Morales M., Ribas I., 2012, ApJ, 753, L25
Rothman L. S. et al., 2010, J. Quant. Spectrosc. Radiat. Transfer, 111, 2139
Scandariato G. et al., 2021, A&A, 646, A159
Schwarz H., Brogi M., De Kok R. J., Birkby J., Snellen I., 2015, A&A, 576, A111
Schwarz H., Ginski C., De Kok R. J., Snellen I. A. G., Brogi M., Birkby J. L., 2016, A&A, 593, A74
Serindag D. B., Nugroho S. K., Mollière P., De Mooij E. J. W., Gibson N. P., Snellen I. A., 2021, A&A, 645, A90
Smette A. et al., 2015, A&A, 576, 77
Snellen I. A., De Kok R. J., De Mooij E. J. W., Albrecht S., 2010, Nature, 465, 1049
Tamuz O., Mazeh T., Zucker S., 2005, MNRAS, 356, 1466
Triaud A. et al., 2018, A&A, 506, 377
Vidal-Madjar A. et al., 2010, A&A, 523, A57
Wang J., Ford E. B., 2002, MNRAS, 418, 1822
Wardenier J. P., Parmentier V., Lee E. K. H., Line M. R., Gharib-Nezhad E., 2021, MNRAS, 506, 1258
Webb R. K., Brogi M., Gandhi S., Line M. R., Birkby J., Chubb K. L., Snellen I. A., Yurchenko S. N., 2020, MNRAS, 494, 108
Wright J. T., Eastman J. D., 2014, PASP, 126, 838
Wyttenbach A., Ehrenreich D., Lovis C., Udry S., Pepe F., 2015, A&A, 557, A62

APPENDIX A: PARAMETER POSTERIOR DISTRIBUTIONS

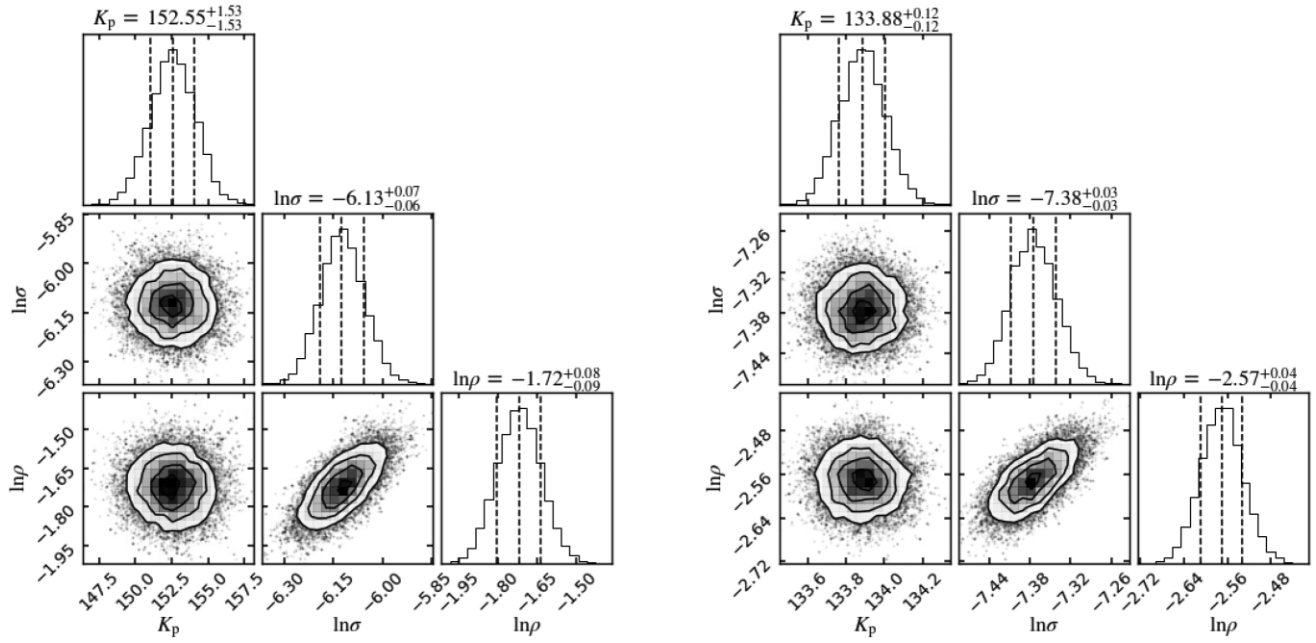

Figure A1. Posterior distributions over radial velocity semi-amplitude K_p and GP hyperparameters from the MCMC, for a GP trained on GP telluric-corrected HD 189733 (*left-hand panel*) and 51 Peg (*right-hand panel*) residuals, with planet spectrum model injections of strengths $n_{inj} = 1.8$ and 9.4, respectively. For each panel, the marginalized distributions for each parameter are given on the diagonal, with dashed lines indicating the locations of the 16th, 50th, and 84th percentiles.

This paper has been typeset from a TEX/LATEX file prepared by the author.