The Astrophysical Journal, 878:153 (11pp), 2019 June 20 https://doi.org/10.3847/1538-4357/ab22b2

A Principal Component Analysis-based Method to Analyze High-resolution Spectroscopic Data on Exoplanets

M. Damiano[1,2], G. Micela[2], and G. Tinetti[1]

[1] Department of Physics & Astronomy, University College London, Gower Street, WC1E 6BT London, UK; mario.damiano.15@ucl.ac.uk
[2] INAF—Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, I-90134 Palermo, Italy

Received 2018 June 19; revised 2019 May 16; accepted 2019 May 18; published 2019 June 25

Abstract

High-resolution spectroscopy has been used to study the composition and dynamics of exoplanetary atmospheres. In particular, the spectrometer CRIRES installed on the ESO-VLT has been used to record high-resolution spectra in the near-IR of gaseous exoplanets. Here we present a new automatic pipeline to analyze CRIRES data sets. Said pipeline is based on a novel use of the principal component analysis and the cross-correlation function. The exoplanetary atmosphere is modeled with the $\mathcal{T}$-REx code using opacities at high temperatures from the ExoMol project. In this work we tested our analysis tools on the detection of CO and H_2O in the atmospheres of the hot Jupiters HD209458b and HD189733b. The results of our pipeline are in agreement with previous results in the literature and other techniques.

Key words: methods: data analysis – planets and satellites: atmospheres – techniques: spectroscopic

1. Introduction

More than 4000 confirmed exoplanets are currently listed in the catalogs, together with basic planetary, stellar, and orbital parameters as they become known. Transit and direct imaging spectroscopy from space and ground facilities are enabling the study of the physical and chemical properties of some of these exoplanets. From space, one can observe exoplanet spectra in the UV, VIS, and IR at low spectral resolution, without the hurdle of telluric contamination. Molecules, ions, atoms, or absorbers able to imprint strong modulations in the recorded spectra can be detected by using space-borne facilities, (e.g., Charbonneau et al. 2002; Tinetti et al. 2007; Grillmair et al. 2008; Linsky et al. 2010; Fraine et al. 2014; Sing et al. 2016; Tsiaras et al. 2016a, 2016b, 2018; Damiano et al. 2017). By contrast, observations from the ground at high-resolution ($R > 25{,}000$) have enabled the detection of molecules or atoms whose weak absorptions are hard to detect at low spectral resolution. This is particularly true for alkali metals and CO that have been found in the atmospheres of most hot Jupiters analyzed (Redfield et al. 2008; Snellen et al. 2010; Birkby et al. 2013, 2017; de Kok et al. 2013; Brogi et al. 2014, 2016; Birkby 2018).

High-resolution spectroscopy (HRS) allows us to resolve molecular bands into individual lines. Using radial velocity measurements and techniques such as the cross-correlation function (CCF) we may separate three physically different sources: telluric absorption, stellar absorption, and the planetary spectrum, which are normally entangled. The aim—but also the biggest challenge—is to recognize the planetary signal among the telluric and the stellar signals, which can be orders of magnitude stronger. The standard method used in the literature to analyze HRS data is to apply a number of corrections that involve the correction of the airmass, the subtraction of a modeled stellar spectrum from the data, and the use of ad hoc masks to eliminate residual strong features (Snellen et al. 2010; Birkby et al. 2013, 2017; Brogi et al. 2014, 2016; Birkby 2018).

In this paper we present and assess an alternative automatic procedure to analyze HRS data from the raw images to the final result, which requires no manual intervention that could interfere with the objectivity and repeatability of the analysis. Our analysis method is based on a novel use of the principal component analysis (PCA) and CCF. The exoplanetary atmosphere has been simulated using $\mathcal{T}$-REx (Waldmann et al. 2015a, 2015b) and line lists have been adopted from the ExoMol project (Tennyson et al. 2016).

We applied our analysis method to two data sets recorded with VLT/CRIRES freely available on the ESO archive. The exoplanets observed are HD209458b and HD189733b (see Table 1), the most studied planets up to date, and therefore good examples for testing new and/or different data analysis techniques. HD209458b (Mazeh et al. 2000) was the first planet analyzed with high-resolution spectroscopy: Snellen et al. (2010) reported a detection of CO in its atmosphere. CO is absent in the Earth's atmosphere but also in the stellar spectrum due to the relatively hot temperature of HD209458. The CO signal in the exoplanetary atmosphere should not be contaminated by the star and Earth's atmosphere. By contrast the star hosting HD189733b is a K type (Bouchy et al. 2005) showing CO absorption features in its spectrum: additional caution is therefore needed to remove the potential stellar contamination. Brogi et al. (2016) have reported the detection of H_2O and CO in the atmosphere of HD189733b.

In Section 2 we describe our analysis method, in Section 3 we show the results, and in Section 4 discussion and conclusions are presented.

2. Data Analysis

We selected data sets relative to HD189733b and HD209458b that are publicly available on the ESO archive. These are part of 289.C-5030(A) and 383.C-0045(A) programs (PI: I. Snellen; Figures 1(a) and 2(a)). The observations have been recorded by using VLT/CRIRES at the highest resolution available ($R = 100{,}000$) through the 0.″2 slit. Both data sets cover a narrow wavelength range, i.e., 2287.54–2345.34 nm and

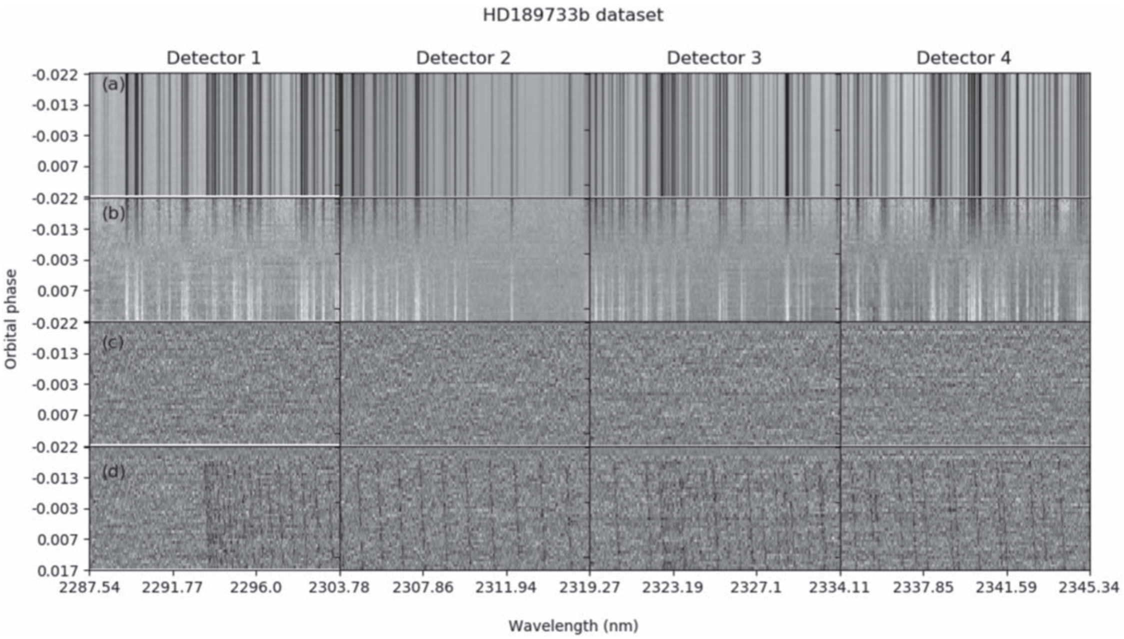

Figure 1. HD189733b data set. In (a), the data are shown after calibration, normalization, and spike correction. In (b), the data are shown after the mean has been subtracted from each column. In (c), the results of PCA are shown. In (d), the data are shown after the application of PCA and the injection of the CO model.

2291.79–2349.25 nm, respectively, with three gaps (∼200 pixels per gap) due to the physical separation of CRIRES's detectors. Both data sets have been recorded with the nodding method ABBA for a better background subtraction (Snellen et al. 2010; Brogi et al. 2016). The steps of the analysis process are represented in Figure 3 and are described in following sections.

Table 1
Relevant Parameters of the Studied Targets

Parameter	HD189733	HD209458
Stellar Parameters		
$R_\star$ ($R_\odot$)	0.756 ± 0.018[a]	$(1.155^{+0.014}_{-0.016})$[a]
$T_{\rm eff}$ (K)	5040 ± 50[a]	6065 ± 50[a]
$M_\star$ ($M_\odot$)	0.806 ± 0.048[a]	1.119 ± 0.033[a]
$\log(g_\star)$ (csg)	4.587 ± 0.015[a]	4.361 ± 0.008[a]
v_{sys} (km s^{-1})	-2.361 ± 0.003[b]	-14.7652 ± 0.0016[c]
Planet Parameters		
$T_{\rm eq}$ (K)	(1201^{+13}_{-12})[a]	1449 ± 12[a]
a (au)	0.03120(27)[d]	$(0.04707^{+0.00046}_{-0.00047})$[a]
R_p ($R_{\rm Jup}$)	$(1.178^{+0.016}_{-0.023})$[d]	$(1.359^{+0.016}_{-0.019})$[a]
M_p ($M_{\rm Jup}$)	$(1.144^{+0.057}_{-0.056})$[a]	0.685 ± 0.015[a]
P (days)	2.21857567(15)[e]	3.52474859(38)[f]
T_0 ($\rm BJD_{UTC}$)	2454279.436714(15)[e]	2452826.629283(87)[f]
I (deg)	85.710 ± 0.024[e]	86.71 ± 0.05[a]

Notes.
[a] Torres et al. (2008).
[b] Bouchy et al. (2005).
[c] Mazeh et al. (2000).
[d] Triaud et al. (2009).
[e] Agol et al. (2010).
[f] Knutson et al. (2007).

2.1. Data Reduction and Calibration

We adopted the pipeline provided by ESO (Crire kit Version-2.3.3) to process the raw data. The CRIRES's reduction pipeline has been embedded into our code thanks to ESO's *EsoRex*, which is a command-line-driven utility that can launch pipeline reduction routines (they are referred to as recipes). These are individual scripts that perform specific actions to the input data. The reduction process performs the following steps:

1. Dark subtraction;
2. Correction for detector nonlinearity;
3. Flat-fielding;
4. Combination of nodding exposures;
5. Spectrum extraction;
6. Wavelength calibration.

The master reduction files (e.g., dark and flat) are provided with the raw data, while the specific nonlinearity correction files need to be downloaded from the archive.[3] The 1D spectrum is extracted from the reduced images via an optimal extraction (Horne 1986). By using the ABBA nodding method, we obtained 45 spectra for the HD189733b data set and 51 for the HD209458b data set.

To subtract and correct the telluric absorption, the calibration from the ESO pipeline is not accurate enough; we followed instead the procedure described in the literature (Snellen et al. 2010; Birkby et al. 2013, 2017; Brogi et al. 2013, 2014, 2016; de Kok et al. 2013), which involves a further calibration using the SKYCALC tool.[4] This simulates the telluric absorption spectrum for a specific night.

The first step is to normalize each spectrum of each detector by dividing it by its median. This step is necessary to avoid differences of baseline across spectra. After the normalization,

[3] https://www.eso.org/sci/facilities/paranal/instruments/crires/doc/VLT-MAN-ESO-14200-4032_v91.pdf

[4] https://www.eso.org/observing/etc/bin/gen/form?INS.MODE=swspectr+INS.NAME=SKYCALC

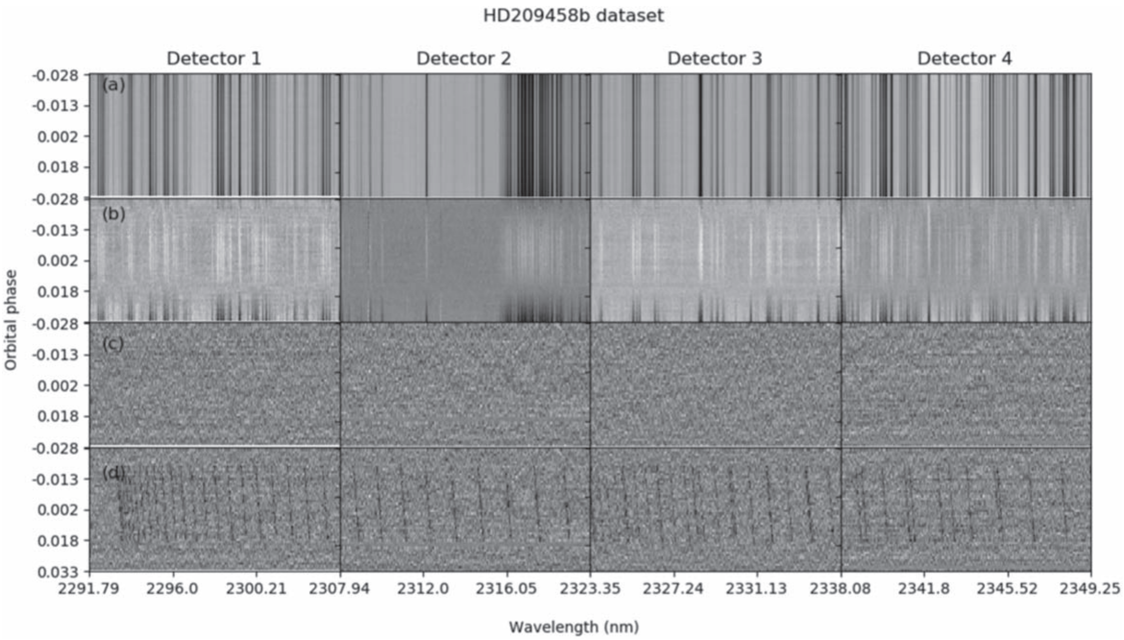

Figure 2. HD209458b data set. In (a), the data are shown after calibration, normalization, and spike correction. In (b), the data are shown after the mean has been subtracted from each column. In (c), the results of PCA are shown. In (d), the data are shown after the application of PCA and the injection of the CO model.

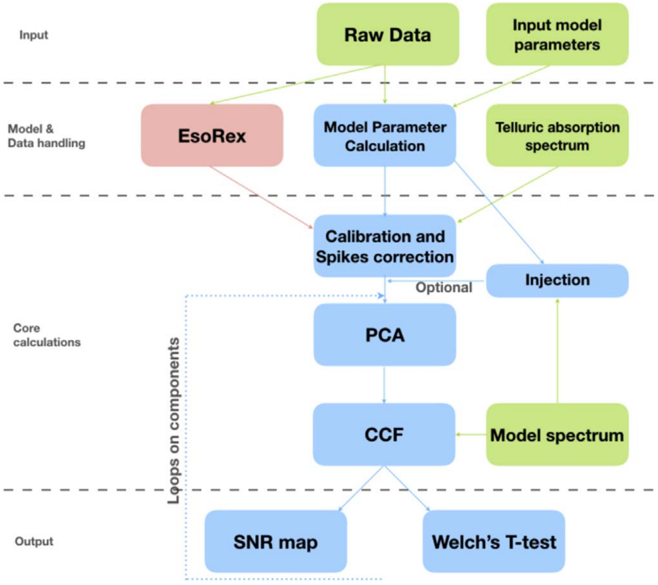

Figure 3. Box colors indicate different classes of action: green boxes represent an external input coming from other models or different sources (e.g., user), the red box includes the external reduction algorithm. Finally, the blue boxes contain the calculations developed for the analysis of the data.

working on one detector at a time, we consider the mean spectrum. Here, the strongest lines (all the lines reaching a minimum <0.8) have been identified as homogeneously distributed as possible to cover the whole x-axis range. These same lines are also been identified within the telluric template. A Gaussian fit is then performed for each of these lines and the centroid is taken. The extracted spectrum centroids indicate the pixel number position. In the telluric template, instead, they indicate wavelength positions of the lines. We performed a fourth-order polynomial fit to establish the relationship between pixels and wavelengths (Snellen et al. 2010). All the single spectra are then interpolated via a third-order spline to the derived wavelength grid to have the same grid for all the spectra.

We analyzed each detector separately as a two-dimensional matrix, where the x-axis contains wavelengths and y-axis time: every row of this matrix is a spectrum, every column is a temporal series at a given wavelength (see Figures 1(a) and 2(a)). We therefore have four different matrices. Finally, the pipeline removes all the cosmic rays or spikes that could occur at the edges of the spectra due to the spline interpolation to the wavelength grid. The pipeline takes one column at a time of each 2D matrix, it calculates the median of the column, and all the values outside 3σ from the median are set to the median value.

2.2. *Decomposition Analysis (PCA)*

The next steps involve the correction for telluric absorption, the subtraction of stellar signal, and the subtraction of correlated noise. The use of an ad hoc mask to remove the strongest telluric features has been frequently adopted in the literature (e.g., Snellen et al. 2010; Brogi et al. 2016). Other works have considered an unsupervised linear transformation technique to identify patterns in data, i.e., PCA. In Artigau et al. (2014) PCA was used to correct high-resolution spectra and improve the radial velocity accuracy for low-mass planetary detection. Similarly, in de Kok et al. (2013), Ridden-Harper et al. (2016), and Piskorz et al. (2016, 2017), PCA has been used to identify and detrend the telluric absorption. In those works PCA was used to decompose the data in the wavelength domain. Here, we explore the use of PCA applied to both wavelength and time domains. Additionally, we propose an objective criterion to determine an optimal selection of the principal components to be considered and the exact number of components to be subtracted. More recently, the algorithm SYSREM developed by Tamuz et al. (2005) has been adopted to perform a similar task (Birkby et al. 2017; Nugroho et al. 2017). SYSREM allows us to extract components iteratively one by one; however, the orthogonality

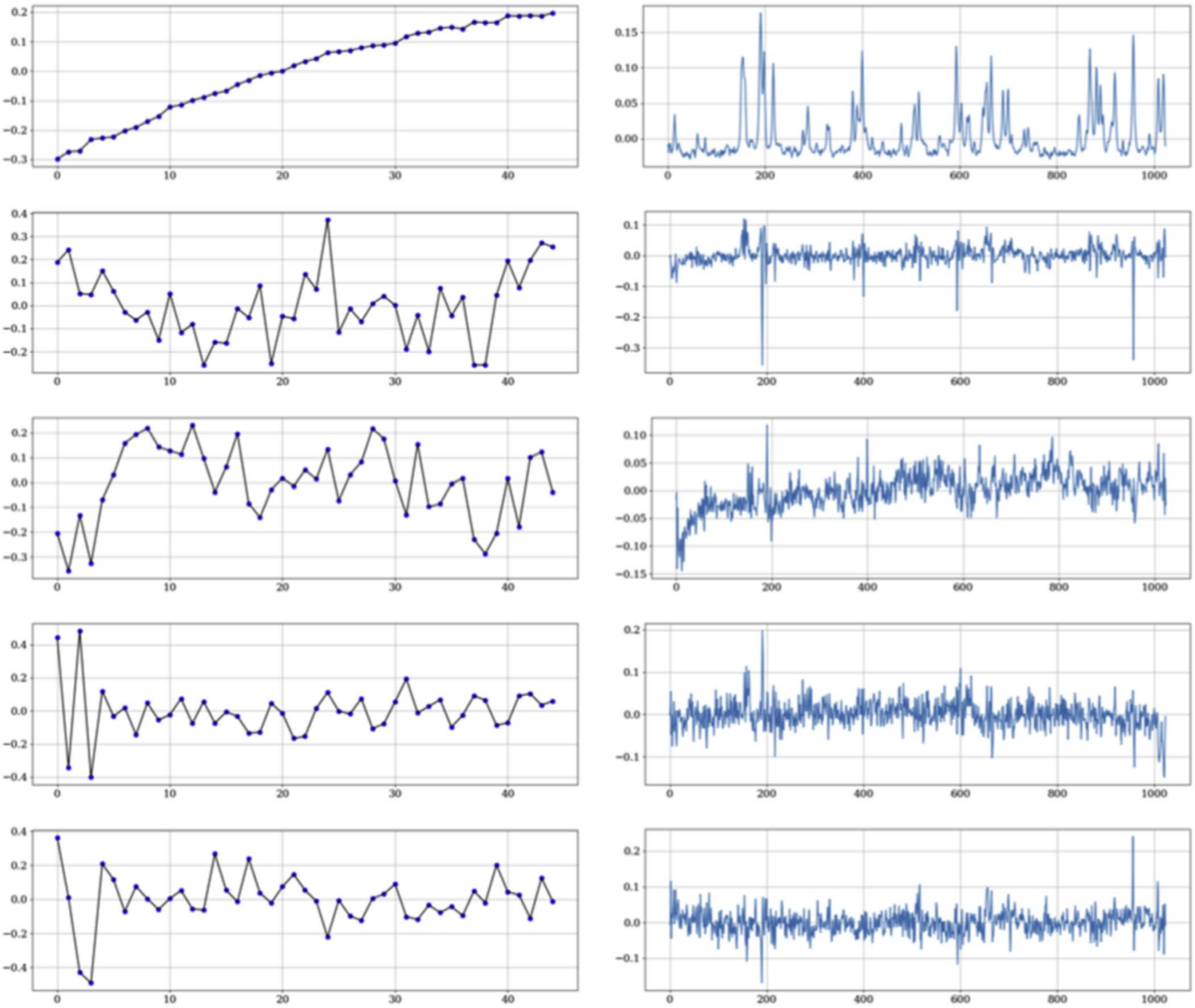

Figure 4. Left panels: first five eigenvectors of the TDM case. Right panels: first five eigenvectors of the WDM covariance matrix.

of the extracted components is not guaranteed (Tamuz et al. 2005).

As PCA is highly sensitive to data scaling, we subtracted each column of the data matrices by its mean (Figures 1(b) and 2(b)). On a typical spectroscopic data set, the number of spectra are less than the wavelength bins, resulting in matrices that have more columns than rows. Here, we adopt the eigenvalue decomposition (EVD) of the covariance matrix (Jolliffe 2002). The dimension of the covariance matrix and the number of principal components (eigenvectors) are equal to the number of rows of the input matrix. Two cases are then considered:

1. *Time domain matrix (TDM)*; we use the individual spectra as rows and the wavelength bins as columns.
2. *Wavelength domain matrix (WDM)*; we transpose the matrix to have the spectra as columns and wavelength bins as rows.

In the WDM/TDM case the principal components (eigenvectors) contain the information of the correlations in the wavelength/time domain. We consider, for example, the first detector of the HD189733b data set: Figure 4 shows the first five components of the TDM case (left) and the first five of the WDM (right). The TDM components contain the time-domain information and the first one, in particular, is linked to the variations of the airmass: these are linearly correlated as we can appreciate from Figure 5. The WDM components show the correlation in the wavelength domain and they appear to be correlated with the telluric transmission spectrum. A good example is the strong feature around 200 (Figure 4, *x*-axis unit, ∼2290 nm) that persists in all the components.

The TDM case has been chosen as best method for the following reasons:

1. The WDM component space cannot be fully described since there are more variables (1024 spectral bins) than observations (45 spectra for HD189733b and 51 for HD209458b data set). The eigenvalues are null after the 44th or 50th component, depending on the data set.
2. The application of a telluric mask is required if the WDM case is chosen to remove most prominent telluric features that persist after PCA has been applied.

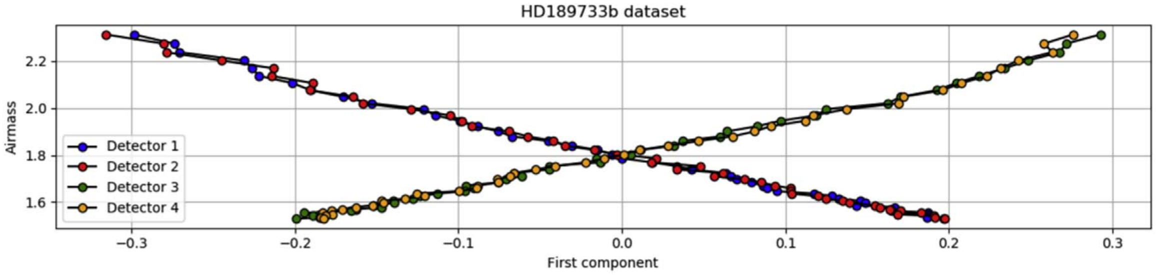

Figure 5. Linear relation between the first component of each detector in the time domain and the recorded airmass for the HD189733b data set.

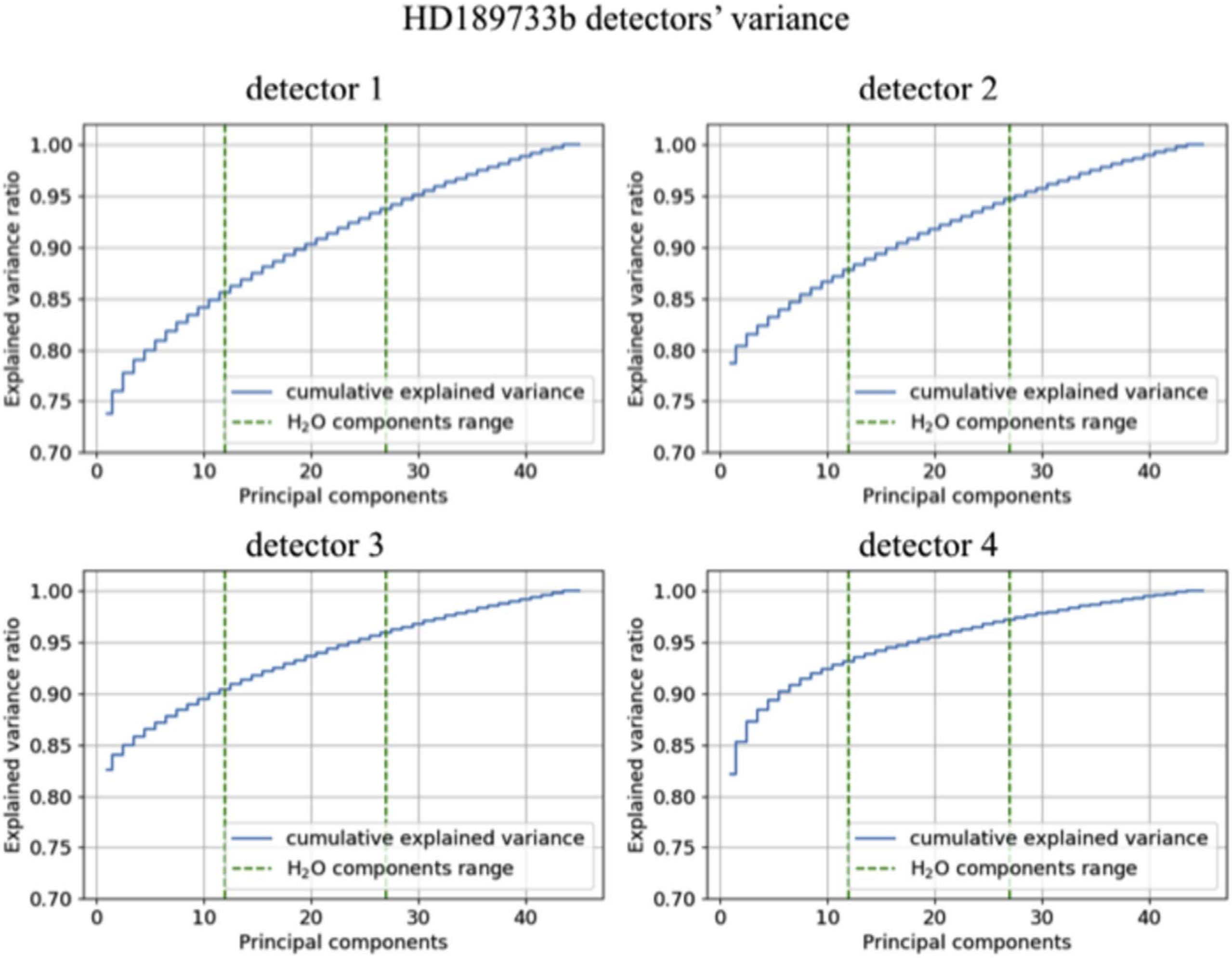

Figure 6. Detectors' variances of the PCA decomposition relative to the HD189733bb data set. The first component always carries more than 75% of the information. However, the variance is different for each of the detectors. The green dashed lines indicate the calculated component range relative to the water vapor.

Following this choice, we calculated 50 TDM components for the HD209458b data set and 44 components for the HD189733b data set. They are equal to the number of recorded spectra minus one, due to the normalization performed before the PCA decomposition. From the eigenvalues we calculated the explained variance ratio (EVR) as follows: $\mathrm{EVR}_j = \lambda_j/\sum\lambda_i$, where λ_i are the eigenvalues. The EVR estimates the information carried by each principal component in percentage. The EVRs of each principal component for every detector of both data sets are shown in Figures 6 and 7. The first component always has the largest variance (∼80%) as the telluric signal is the most significant.

The components are chosen to maximize the signal-to-noise ratio (S/N) of the CCF peak expected at the theoretical K_p and v_{rest} of the planet (see Equation (3); Section 2.5; Figures 6 and 7). This task is accomplished by removing iteratively the low-order components, which supposedly are telluric or stellar in origin, and the high-order components, which account for noncorrelated signal, presumably noise. The remaining components (time domain eigenvectors) are then projected back onto the original space.

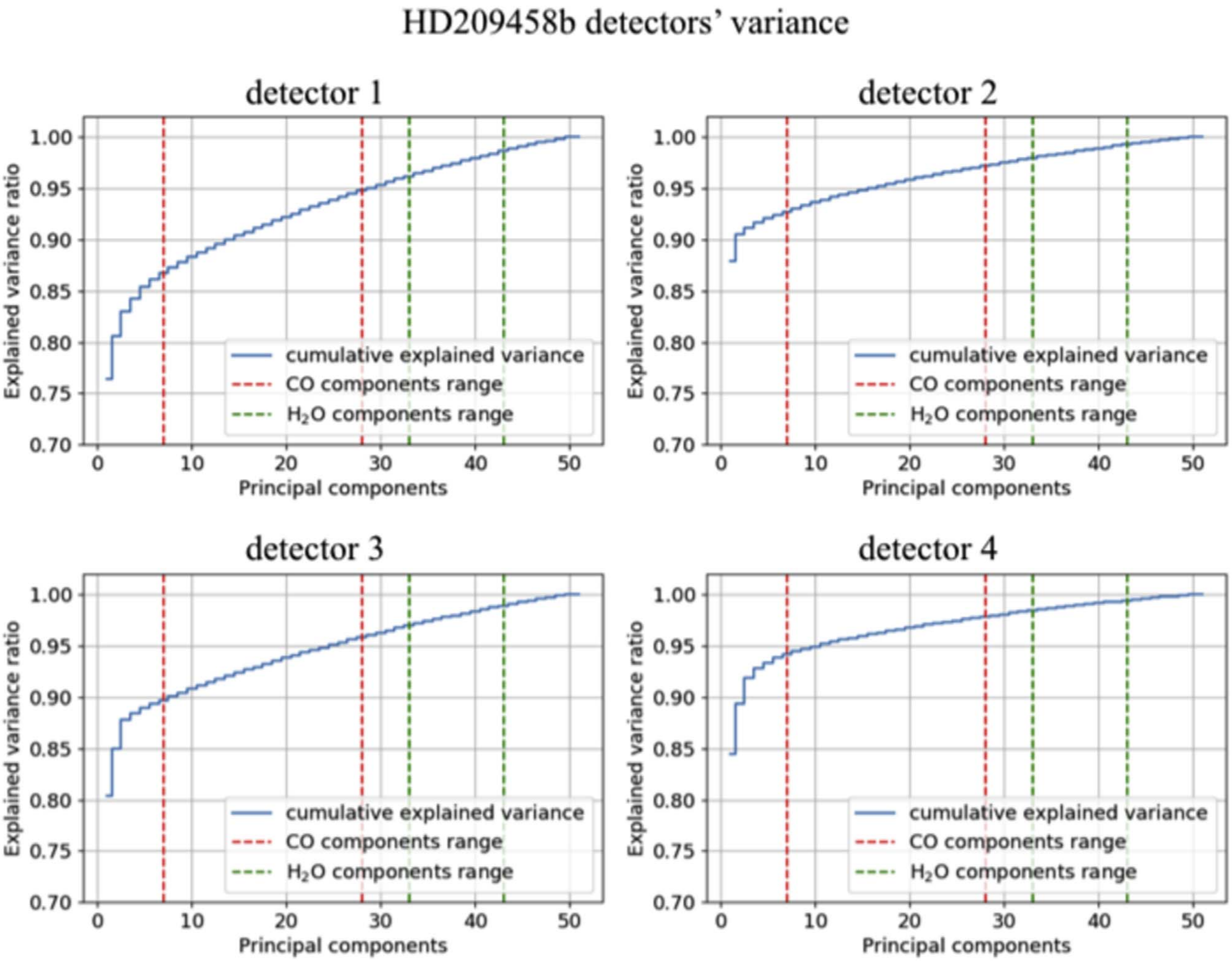

Figure 7. Detectors' variances of the PCA decomposition relative to the HD209458b data set. The first component always carries more than 75% of the information. However, the variance is different for each of the detectors. The red dashed lines highlight the determined component range relative to the CO, while the green dashed lines are relative to the H_2O.

After the application of PCA, each column of the output matrix was divided by its standard deviation to restore the S/N of the processed data (Birkby et al. 2013; de Kok et al. 2013; Nugroho et al. 2017; Ridden-Harper et al. 2016; Figures 1(c) and 2(c)).

2.3. Cross-correlation Function

The CCF measures the similarity of two signals. It is also often called *sliding dot product* since it returns a single value from the product of two signals when one slides over the other. Considering two series $\boldsymbol{x}$ and $\boldsymbol{y}$, the normalized CCF at the delay d, for discrete series, is defined as follows (Bracewell 1965):

$$\mathrm{CCF}(d) = \frac{\sum_i ((x(i) - \bar{x}) \cdot (y(i-d) - \bar{y}))}{\sqrt{\sum_i (x(i) - \bar{x})^2} \cdot \sqrt{\sum_i (y(i-d) - \bar{y})^2}}, \quad (1)$$

where $\bar{x}$ is the mean of the array $\boldsymbol{x}$, $\bar{y}$ is the mean of the array $\boldsymbol{y}$, and $i = 0, 1, 2 \ldots N - 1$. The idea of using such a function is to find possible correlations between the data and an atmospheric model. The cross-correlation aims at matching similarities between the two signals.

The exoplanet atmospheric models have been simulated using $\mathcal{T}$-REx (Waldmann et al. 2015a, 2015b). The CO and H_2O line lists at the planetary temperature were provided by ExoMol (Tennyson & Yurchenko 2012; Tennyson et al. 2016).

Every row of the data matrix (every single spectrum), after the application of PCA, is cross-correlated with the simulated exoplanet atmospheric spectrum. This spectrum is interpolated to the same wavelength grid of the data, and it is then shifted from -100 to 100 km s^{-1} with 1.0 km s^{-1} as the step. The step is chosen based on the precision obtained during the calibration step ($\sim$1.0 km s^{-1}) and on the velocity resolution of the instrument (1.5 km s^{-1}).

The CCF transforms the matrices (one for each of the four detectors of CRIRES) from the wavelength domain to the velocity domain. The CCF matrices are then added together to obtain one single matrix (we will refer to it as the CCF matrix).

At this stage the exoplanetary signal is not visible (see Figure 8, top left panel). We then injected a synthetic signal with the orbital parameters of the planet to predict the position of the signal (Figure 8, bottom left panel) and to calculate the area of the S/N matrix interested by the planetary signal (see

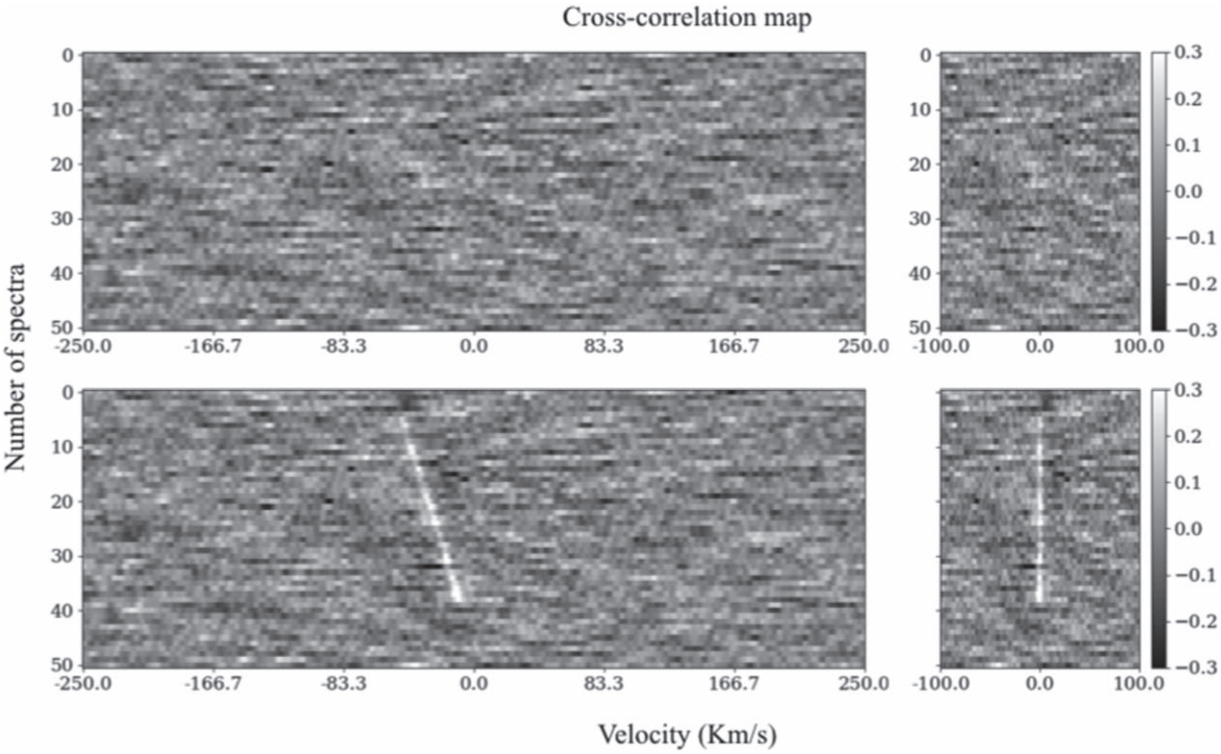

Figure 8. Top left panel: the four CCFs of the four detectors of CRIRES of the HD209458b data set summed together. Bottom left panel: same as top but with the model injected. The injection is 1× the synthesized model ($R_p/R_\star \sim 10^{-3}$). Top right panel: cross-correlation after changing the reference frame from the Earth to the rest frame of the exoplanet. In this frame the planetary cross-correlation signal is aligned to 0 km s^{-1}. Bottom right panel: same as the top right panel, but with the injection. The injected signal is aligned to 0 km s^{-1} in the exoplanet's rest frame.

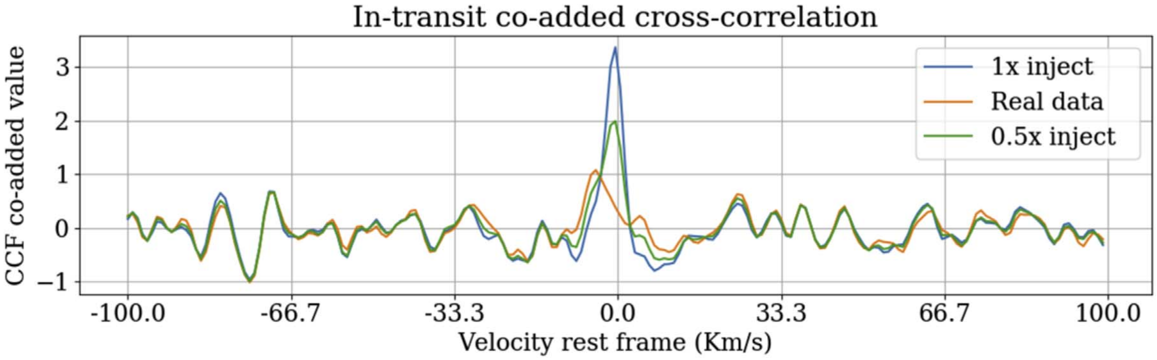

Figure 9. Cross-correlations of water vapor coadded in transit for the HD209458b data set. The injected signal and the planetary signal are still present after using PCA. The coadded CCFs are relative to the HD209458b rest frame ($K_p = 145.041$ km s^{-1}). This graph has been generated considering PCA components from 33 to 43.

Section 2.5 and Figures 9 and 10). This step is performed after the calibration (before the PCA decomposition) and the effects of this process are not visible until the CCF is performed, because the signal intensity is at least three orders of magnitude weaker than the telluric and stellar signals. The injected model cross-correlates with itself, resulting, for example, in the signal shown in Figure 8 (bottom panels). Finally, the injection process allows us to monitor the signal during the PCA decomposition. That helps to determine when the component subtraction starts to erase part of the signal.

2.4. Signal Extraction

At this stage the planetary signal was barely visible or completely invisible, therefore we coadded the single CCFs (rows of the CCF matrix) in transit to obtain the integrated signal from the planet. As the data were aligned to the telluric spectrum reference system, the planetary spectrum moved across time; we then realigned the single CCFs to the reference system of the planet by computing the following correction:

$$V_p = K_p \sin[2\pi\phi(t)] + v_{\rm sys} + v_{\rm bary}(t), \tag{2}$$

where K_p is the radial velocity amplitude of the planet (Equation (3)) and $\phi(t)$ is the orbital phase (Equation (5)):

$$K_p = v_{\rm orb} \sin(i), \tag{3}$$

$$v_{\rm orb} = \frac{2\pi a}{P_{\rm orb}}, \tag{4}$$

$$\phi(t) = \frac{t - T_0}{P_{\rm orb}}. \tag{5}$$

All parameters are listed in Table 1.

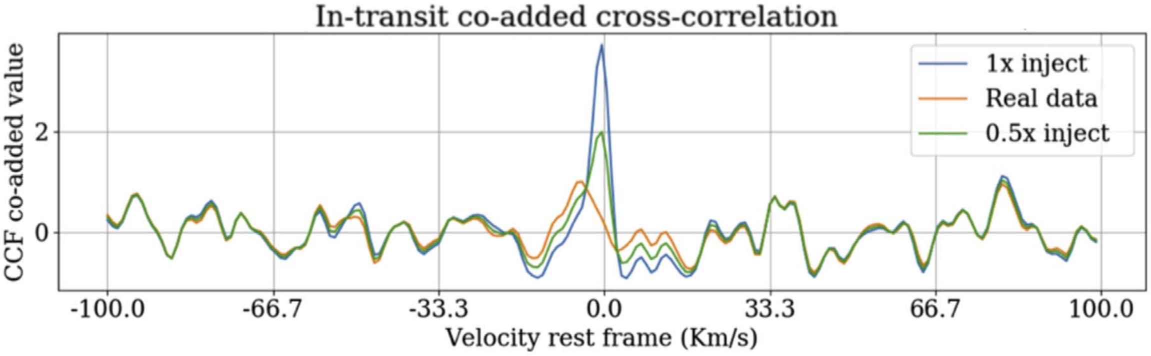

Figure 10. Cross-correlations of the planetary signal and of the injected water vapor coadded in transit for the HD189733b data set. The coadded CCFs are calculated at the theoretical orbital velocity of the planet HD189733b ($K_p = 152.564$ km s^{-1}). The CCFs are the result of the combination of the PCA components from 12th to 27th.

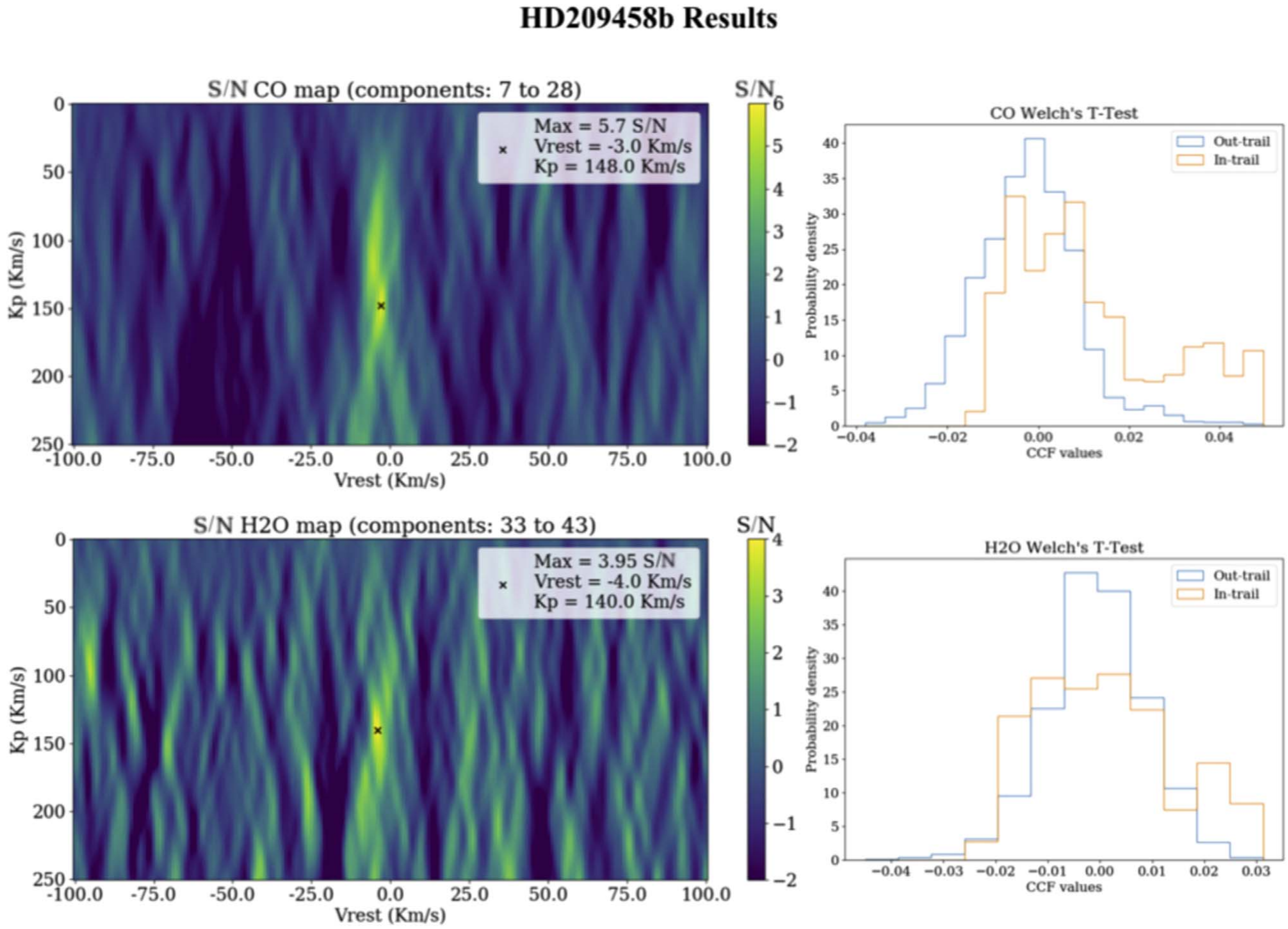

Figure 11. Results for the HD209458b data set. Top left panel: S/N map for the carbon monoxide. The maximum point is compatible with the planetary orbital parameters. Top right panel: distributions (i.e., in-trail and out-trail) used to compute the Welch's T-test. The null hypothesis is rejected with a confidence greater than 7σ. Bottom left panel: S/N map of the water vapor. The peak is compatible with the planetary parameters. Bottom right panel: distribution used to compute the Welch's T-test. The null hypothesis is rejected with a confidence of 6.56σ.

Once all CCFs were aligned to the planetary rest frame, we coadded in time only the in-transit CCFs. These were selected by computing the transit time (Seager & Mallén-Ornelas 2003; Kipping 2010). When all the in-transit cross-correlations are summed together, the 2D cross-correlation matrix is reduced to a 1D signal, which is connected to the theoretical orbital velocity of the planet. To explore different orbital velocities we proceeded as follows:

1. We let K_p vary from 0 to 250 km s^{-1} with a 1 km s^{-1} step;
2. For each K_p we applied the correction in Equation (2) to every single CCF in the CCF matrix; and
3. We summed only the in-transit cross-correlations.

In this way we were able to explore all possible orbital velocities including those corresponding to the host star. Following the previous steps, we obtained a matrix with K_p on

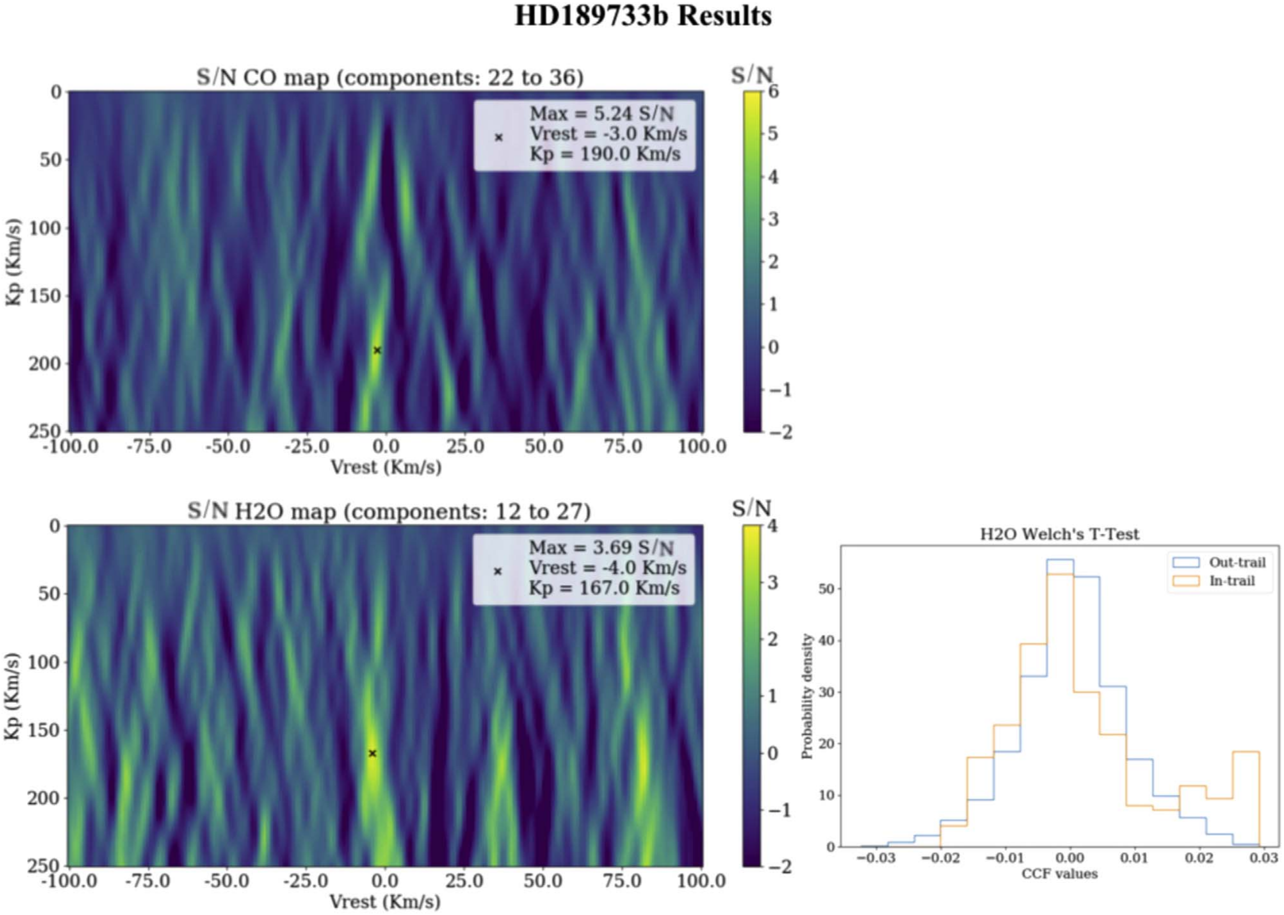

Figure 12. Results for the HD189733b data set. Top left panel: S/N map for the carbon monoxide. The maximum point is compatible with the result reported in Brogi et al. (2016) but it is not compatible with the expected value. Bottom left panel: S/N map of the water vapor. The peak is compatible with the planetary parameters. Bottom right panel: distribution used to compute the Welch's T-test. The null hypothesis is rejected with a confidence of 5.21σ.

the y-axis and the velocity rest frame along the x-axis (v_{rest}). From this matrix two different outputs were extracted: the S/N map and the T-test statistic.

2.5. S/N Matrix

We considered the last matrix obtained, i.e., K_p on the y-axis and v_{rest} on the x-axis. We calculated the standard deviation of this matrix excluding those points potentially correlated to the planetary signal ($|v_{\mathrm{rest}}| < 15\,\mathrm{km\,s^{-1}}$), and we divided the entire matrix by this value. We refer to the obtained matrix as the S/N matrix (Figure 11, left panels and Figure 12, left panels).

To assign an uncertainty to the K_p value we followed the same procedure as reported in Brogi et al. (2016), i.e., we took the maximum value of the matrix and, fixing the relative v_{rest}, we calculated the K_p interval where the S/N dropped by a unit around the K_p peak. The same approach was used to determine the uncertainty for v_{rest}.

The S/N map is not only useful to visually represent the results but also to inspect whether spurious signals or telluric residuals are present. These signals may have a high-S/N value but are located at different K_p and/or v_{rest} from those expected for the planetary signal.

We calculated the S/N matrix for each excluded principal component. Two loops need to be performed to explore the entire principal component space: the first loop subtracts higher variance components onward and aims to remove the most correlated signal (e.g., telluric absorption and stellar signal). The second loop subtracts lower variance components backward and aims to remove uncorrelated noise from the data. Finally, the principal components were selected to maximize the peak of the S/N matrix in correspondence with the expected planetary K_p and v_{rest}.

2.6. Welch's T-test Statistics

The Welch's T-test is used to test the hypothesis that two populations have equal means. This test compares the population of points on the CCF map connected to the planetary signal with those that are not.

From the CCF matrix we defined, as done in the literature (Brogi et al. 2016; Nugroho et al. 2017):

1. In-trail, those values inside a squared box centered on the CCF' peak with a radius of $\pm 15\,\mathrm{km\,s^{-1}}$;
2. Out-trail, those values outside the in-trail box.

We extracted two families of values from the CCF matrix and these were compared through the Welch's T-test (Figure 11, right panels and Figure 12, right panels). The test, calculated using *scipy.stats.ttest ind* in *python*, provides a p-value (two-tailed) that was converted into σ value (significance interval) through the inversion of the survival function (SF):

$$\sigma_{\mathrm{value}} = \mathrm{SF}^{-1}(p\text{-value} \;/\; 2), \tag{6}$$

where the SF^{-1} is the inverse of the SF that is calculated from the cumulative density function (CDF) as follows:

$$SF = 1 - CDF. \qquad (7)$$

3. Results

HD209458b: For the cross-correlation process, we assumed an isothermal $T - p$ profile at $T = 1400$ K, with the pressure varying from 10^{-5} to 10^4 Pa. We did not include clouds or line broadening due to the rotation of the planet. We used 10^{-3} as volume mixing ratio (VMR) for both molecules; this value is compatible with chemical models' predictions for hot-Jupiter atmospheres (Venot et al. 2012). The same value was also used by Snellen et al. (2010) for the CO.

The signal obtained for CO peaks at S/N = 5.7 (Figure 11, top left panel and Table 2). The signal is compatible with the planetary orbital parameters ($K_p = 148^{+16}_{-15}$ km s^{-1}, $v_{\rm rest} = -3.0^{+1.3}_{-1.1}$ km s^{-1}). This result has been obtained by considering components from the 7th to the 28th (Figure 7, red lines). The statistical significance of the result is also confirmed by the Welch's T-test (Figure 11, top right panel). Using a box of radius 15 km s^{-1} the null hypothesis is rejected with a confidence greater than 7σ; the shift of the in-trail population is noticeable with respect to the out-trail values that are, instead, distributed as a Gaussian centered to zero.

The signal of the water vapor is more difficult to detect since the Earth's atmosphere also contains water. To extract the planetary signal a robust telluric correction is required, and therefore several components need to be subtracted using PCA. A signal at the compatible planetary parameters is observable in the S/N map in Figure 11 (bottom left panel). The maximum peaks at S/N = 3.95, $K_p = 140^{+25}_{-16}$ km s^{-1}, and $v_{\rm rest} = -4.0^{+1.4}_{-1.6}$ km s^{-1} and it is obtained considering components from the 33th to the 43th (Figure 7, green lines). To demonstrate that the H_2O planetary signal survives after 33 components have been subtracted, Figure 9 shows the in-transit coadded cross-correlation relative to the range of components aforementioned. Both the injected and non-injected signals survive to the PCA correction (note that the injected signal does not include any atmospheric dynamics, so it is not blueshifted like the planetary signal). Moreover, the coadded cross-correlation value is lower with respect to the CO case meaning that the concentration of water is lower than CO or that PCA has erased part of the signal. Finally, the Welch's T-test is performed on the in-trail and out-trail populations (Figure 11, bottom right panel and Table 2). In this case the shift of the in-trail population is not as strong as in the CO case but the null hypothesis is rejected with a confidence greater than 6σ.

HD189733b: The planetary transmission spectrum was modeled with isothermal $T - p$ profiles at $T = 1000$ K. The pressure varies from 10^{-5} to 10^4 Pa, and we did not include clouds or any line broadening due to the rotation of the planet. We used 10^{-3} as the VMR; this value is compatible with chemical model predictions for hot Jupiters (Venot et al. 2012).

CO detection is highly difficult since the star, being a K-type star ($T \sim 4900$ K), contains CO in the outer regions. In Brogi et al. (2016) a master stellar spectrum has been simulated and subtracted to the data, but the stellar contamination continued to be persistent also in the result. In this work, PCA was not as effective as in the HD209458b case because the star spectrum moves 1–2 pixels on the detector preventing an optimal correction. The result (see Figure 12, top left panel and Table 2) is compatible with the one claimed by Brogi et al. (2016) (S/N = 5.1, $K_p = 194^{+19}_{-41}$ km s^{-1}, $v_{\rm rest} = -1.7^{1.1}_{1.2}$ km s^{-1}); however, the error on the K_p, being smaller than the one reported in literature, does not include the theoretical value of the orbital velocity of the planet ($K_p = 152.564$ km s^{-1}). The signal determined at lower K_p ($\sim$85 km s^{-1}) is due to stellar contamination that results from the Rossiter–McLaughlin effect combined with the change of reference frame from the Earth to the barycentric one (Brogi et al. 2016).

Table 2
This Work and Previous Results

Parameter	HD189733	HD209458
Previous Results	Brogi et al. (2016)	Snellen et al. (2010)
$S/N_{\rm CO}$	…	…
$K_{p,\rm CO}$ (km s^{-1})	205^{+38}_{-51}	…
$v_{p,\rm CO}$ (km s^{-1})	…	140 ± 10
$v_{\rm rest,\ CO}$(km s^{-1})	$-1.6^{+2.0}_{-1.8}$	~2
S/N $_{H_2O}$	5.5	…
K_{p,H_2O} (km s^{-1})	183^{+38}_{-59}	…
$v_{{\rm rest},H_2O}$(km s^{-1})	$-1.58^{+1.65}_{-1.50}$	…
Results	This Work	This Work
$S/N_{\rm CO}$	5.24	5.7
$K_{p,\rm CO}$ (km s^{-1})	190 ± 16	148^{+16}_{-15}
$v_{\rm rest,\ CO}$(km s^{-1})	$-3.0^{+1.0}_{-1.3}$	$-3.0^{+1.3}_{-1.1}$
W T-test$_{\rm CO}$ (σ)	…	21.62
S/N $_{H_2O}$	3.69	3.95
K_{p,H_2O} (km s^{-1})	167^{+32}_{-21}	140^{+25}_{-16}
$v_{{\rm rest},H_2O}$(km s^{-1})	$-4.0^{+2.0}_{-1.8}$	$-4.0^{+1.4}_{-1.6}$
W T-test $_{H_2O}$ (σ)	5.21	6.56

Concerning water vapor, the same discussion as for HD209458b can be applied here. The planetary water signal needs to be disentangled from the telluric absorption. The result obtained (Figure 6, green lines and Figure 12) is compatible with both the literature and the theoretical parameters, e.g., see Figure 10, where the planetary signal is compared with the injected one. The injected signals do not account for $v_{\rm rest} \neq 0$ km s^{-1}; we can appreciate the data being blueshifted. Here the Welch's T-test confirms that the null hypothesis can be rejected with a confidence greater than 5σ (Figure 12, bottom right panel and Table 2).

We performed an additional test by cross-correlating the telluric model used in the calibration process with the data to check if any telluric signal still persists. Using the components reported in the results we did not notice any significant correlation with the telluric signal at the position of the planetary parameters. We have also tried to cross-correlate other molecules with the data (e.g., CH_4, NH_3, and CO_2), but no correlations have been found.

4. Discussion and Conclusions

We presented here and tested a new automatic method, from the raw images to the final result, based on the iterative use of

PCA and CCF to reanalyze two CRIRES data sets observed with a high-resolution spectroscopy technique. Our pipeline does not assume prior knowledge, e.g., the variation of the airmass, nor does it require ad hoc corrections, e.g., masks to remove telluric lines. The PCA components are automatically selected by maximizing the signal extracted. The algorithm is able to calculate the final result (S/N maps and Welch's T-test) without manual intervention, allowing us to analyze rapidly many data sets.

CO and H_2O have been detected in the HD209458b data set, and H_2O in the HD189733b data set. The detection of CO in the HD209458b atmosphere is supported by an S/N peak of 5.7 at K_p and $v_{\rm rest}$ compatible with the planetary orbital parameters. Contrary to CO, H_2O is present in the Earth's atmosphere and therefore an accurate telluric correction is required. The lower S/N peak may be due to a lower concentration of H_2O with respect to CO in the atmosphere of HD209458b, or part of the signal might have been removed by PCA. In both detections a blueshift has been observed and this could be explained with high-altitude winds. The results presented here are in agreement with the results published by Snellen et al. (2010).

Concerning HD189733b, using our method, we have been able to detect H_2O. Even in this data set a blueshift of the signal has been observed and also in this case it could be associated with high-altitude winds. The detected CO signal is compatible with the literature (Brogi et al. 2016), but it is not in agreement with the theoretical radial velocity of the planet, and could be due to stellar contamination (the K-type star shows CO spectral features).

We note that the requirement on maximization of the S/N peak may lead to biased K_p and $v_{\rm rest}$ values. In the work presented here this effect, if present, does not exceed the reported error bars: changes of less than one pixel are found between one component and the others (one pixel corresponds to the CCF step).

We note that the EVR is different for each detector (Figures 7 and 6), and this means that the planetary signal is contained in different components in each detector. An optimal approach should adapt the number of components per detector based on their variance.

Future work will consider the use of the algorithm presented here to analyze high-resolution observations taken by other instruments. These include CRIRES+ (Follert et al. 2014); GIANO-B, a high-dispersion spectrograph at TNG (Oliva et al. 2012), which covers 0.9–2.5 μm with a resolution of ($R = 50{,}000$); IRCS-SUBARU (Kobayashi et al. 2000), which uses a lower resolution ($R = 20{,}000$) but covers a broader range (from 1 to 5μm); and CARMENES at Calar Alto Observatory (Quirrenbach et al. 2014) with a spectral resolution up to 80,000 in the near-IR (0.9–1.7 μm).

M.D. and G.T. are supported by the European Research Council (ERC) grant 617119 ExoLights. M.D. and G.M. are supported by the project "A Way to Other Worlds" funded by Italian Minister for University and Research. This work was also supported by STFC (ST/P000282/1). The authors wish to thank Dr. Antonino Petralia, Dr. Ingo Waldmann, Prof. Jonathan Tennyson, and Dr. Sergey Yurchenko for their input. Finally, we thank the anonymous referee for the comments provided that helped to improve the quality of the manuscript.

ORCID iDs

M. Damiano https://orcid.org/0000-0002-1830-8260
G. Micela https://orcid.org/0000-0002-9900-4751
G. Tinetti https://orcid.org/0000-0001-6058-6654

References

Agol, E., Cowan, N. B., Knutson, H. A., et al. 2010, ApJ, 721, 1861
Artigau, É., Astudillo-Defru, N., Delfosse, X., et al. 2014, Proc. SPIE, 9149, 914905
Birkby, J. L. 2018, arXiv:1806.04617
Birkby, J. L., de Kok, R. J., Brogi, M., et al. 2013, MNRAS, 436, L35
Birkby, J. L., de Kok, R. J., Brogi, M., Schwarz, H., & Snellen, I. A. G. 2017, AJ, 153, 138
Bouchy, F., Udry, S., Mayor, M., et al. 2005, A&A, 444, L15
Bracewell, R. 1965, The Fourier Transform and Its Applications (New York: McGraw-Hill)
Brogi, M., de Kok, R. J., Albrecht, S., et al. 2016, ApJ, 817, 106
Brogi, M., de Kok, R. J., Birkby, J. L., Schwarz, H., & Snellen, I. A. G. 2014, A&A, 565, A124
Brogi, M., Snellen, I. A. G., de Kok, R. J., et al. 2013, ApJ, 767, 27
Charbonneau, D., Brown, T. M., Noyes, R. W., & Gilliland, R. L. 2002, ApJ, 568, 377
Damiano, M., Morello, G., Tsiaras, A., Zingales, T., & Tinetti, G. 2017, AJ, 154, 39
de Kok, R. J., Brogi, M., Snellen, I. A. G., et al. 2013, A&A, 554, A82
Follert, R., Dorn, R. J., Oliva, E., et al. 2014, Proc. SPIE, 9147, 914719
Fraine, J., Deming, D., Benneke, B., et al. 2014, Natur, 513, 526
Grillmair, C. J., Burrows, A., Charbonneau, D., et al. 2008, Natur, 456, 767
Horne, K. 1986, PASP, 98, 609
Jolliffe, I. T. 2002, Principal Component Analysis (Berlin: Springer)
Kipping, D. M. 2010, MNRAS, 407, 301
Knutson, H. A., Charbonneau, D., Noyes, R. W., Brown, T. M., & Gilliland, R. L. 2007, ApJ, 655, 564
Kobayashi, N., Tokunaga, A. T., Terada, H., et al. 2000, Proc. SPIE, 4008, 1056
Linsky, J. L., Yang, H., France, K., et al. 2010, ApJ, 717, 1291
Mazeh, T., Naef, D., Torres, G., et al. 2000, ApJL, 532, L55
Nugroho, S. K., Kawahara, H., Masuda, K., et al. 2017, AJ, 154, 221
Oliva, E., Origlia, L., Maiolino, R., et al. 2012, Proc. SPIE, 8446, 84463T
Piskorz, D., Benneke, B., Crockett, N. R., et al. 2016, ApJ, 832, 131
Piskorz, D., Benneke, B., Crockett, N. R., et al. 2017, AJ, 154, 78
Quirrenbach, A., Amado, P. J., Caballero, J. A., et al. 2014, Proc. SPIE, 9147, 91471F
Redfield, S., Endl, M., Cochran, W. D., & Koesterke, L. 2008, ApJL, 673, L87
Ridden-Harper, A. R., Snellen, I. A. G., Keller, C. U., et al. 2016, A&A, 593, A129
Seager, S., & Mallén-Ornelas, G. 2003, ApJ, 585, 1038
Sing, D. K., Fortney, J. J., Nikolov, N., et al. 2016, Natur, 529, 59
Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W., & Albrecht, S. 2010, Natur, 465, 1049
Tamuz, O., Mazeh, T., & Zucker, S. 2005, MNRAS, 356, 1466
Tennyson, J., & Yurchenko, S. N. 2012, MNRAS, 425, 21
Tennyson, J., Yurchenko, S. N., Al-Refaie, A. F., et al. 2016, JMoSp, 327, 73
Tinetti, G., Vidal-Madjar, A., Liang, M.-C., et al. 2007, Natur, 448, 169
Torres, G., Winn, J. N., & Holman, M. J. 2008, ApJ, 677, 1324
Triaud, A. H. M. J., Queloz, D., Bouchy, F., et al. 2009, A&A, 506, 377
Tsiaras, A., Rocchetto, M., Waldmann, I. P., et al. 2016b, ApJ, 820, 99
Tsiaras, A., Waldmann, I. P., Rocchetto, M., et al. 2016a, ApJ, 832, 202
Tsiaras, A., Waldmann, I. P., Zingales, T., et al. 2018, AJ, 155, 156
Venot, O., Hébrard, E., Agúndez, M., et al. 2012, A&A, 546, A43
Waldmann, I. P., Rocchetto, M., Tinetti, G., et al. 2015a, ApJ, 813, 13
Waldmann, I. P., Tinetti, G., Rocchetto, M., et al. 2015b, ApJ, 802, 107