

## Investigating stellar activity through eight years of Sun-as-a-star observations

Klein, Baptiste; Aigrain, Suzanne; Cretignier, Michael; Al Moulla, Khaled; Dumusque, Xavier; Barragán, Oscar; Yu, Haochuan; Mortier, Annelies; Rescigno, Federica; Cameron, Andrew Collier; López-Morales, Mercedes; Meunier, Nadège; Sozzetti, Alessandro; O'Sullivan, Niamh K.

DOI:  
[10.1093/mnras/stae1313](https://doi.org/10.1093/mnras/stae1313)

License:  
Creative Commons: Attribution (CC BY)

*Document Version*  
Publisher's PDF, also known as Version of record

*Citation for published version (Harvard):*  
Klein, B, Aigrain, S, Cretignier, M, Al Moulla, K, Dumusque, X, Barragán, O, Yu, H, Mortier, A, Rescigno, F, Cameron, AC, López-Morales, M, Meunier, N, Sozzetti, A & O'Sullivan, NK 2024, 'Investigating stellar activity through eight years of Sun-as-a-star observations', *Monthly Notices of the Royal Astronomical Society*, vol. 531, no. 4, pp. 4238-4262. <https://doi.org/10.1093/mnras/stae1313>

[Link to publication on Research at Birmingham portal](#)

### General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.










When citing, please reference the published version.

### Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.

# Investigating stellar activity through eight years of Sun-as-a-star observations

Baptiste Klein <sup>1</sup>★, Suzanne Aigrain,<sup>1</sup> Michael Cretignier <sup>1</sup>, Khaled Al Moulla <sup>2</sup>, Xavier Dumusque,<sup>2</sup> Oscar Barragán <sup>1</sup>, Haochuan Yu <sup>1</sup>, Annelies Mortier <sup>3</sup>, Federica Rescigno <sup>4</sup>, Andrew Collier Cameron <sup>5</sup>, Mercedes López-Morales,<sup>6</sup> Nadège Meunier,<sup>7</sup> Alessandro Sozzetti <sup>8</sup> and Niamh K. O’Sullivan<sup>1</sup>

<sup>1</sup>Department of Physics, University of Oxford, Oxford OX13RH, UK

<sup>2</sup>Observatoire Astronomique de l’Université de Genève, Chemin Pegasi 51, Versoix 1290, Switzerland

<sup>3</sup>School of Physics & Astronomy, University of Birmingham, Edgbaston, Birmingham B152TT, UK

<sup>4</sup>Department of Astrophysics, University of Exeter, Stocker Rd, Exeter EX44QL, UK

<sup>5</sup>Centre for Exoplanet Science / SUPA, School of Physics & Astronomy, University of St Andrews, North Haugh ST ANDREWS, Fife KY16 9SS, UK

<sup>6</sup>Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge MA 02138, USA

<sup>7</sup>Univ. Grenoble Alpes, CNRS, IPAG, Grenoble F-38000, France

<sup>8</sup>INAF – Osservatorio Astrofisico di Torino, via Osservatorio 20, Pino Torinese 10025, Italy

Accepted 2024 May 20. Received 2024 May 20; in original form 2024 February 28

## ABSTRACT

Stellar magnetic activity induces both distortions and Doppler-shifts in the absorption line profiles of Sun-like stars. Those effects produce apparent radial velocity (RV) signals which greatly hamper the search for potentially habitable, Earth-like planets. In this work, we investigate these distortions in the Sun using cross-correlation functions (CCFs), derived from intensive monitoring with the high-precision spectrograph HARPS-N. We show that the RV signal arising from line-shape variations on time-scales associated with the Sun’s rotation and activity cycle can be robustly extracted from the data, reducing the RV dispersion by half. Once these have been corrected, activity-induced Doppler-shifts remain, that are modulated at the solar rotation period, and that are most effectively modelled in the time domain, using Gaussian processes (GPs). Planet signatures are still best retrieved with multidimensional GPs, when activity is jointly modelled from the raw RVs and indicators of the line width or of the Ca II H & K emission. After GP modelling, the residual RVs exhibit a dispersion of 0.6–0.8 m s<sup>−1</sup>, likely to be dominated by signals induced by supergranulation. Finally, we find that the statistical properties of the RVs evolve significantly over time, and that this evolution is primarily driven by sunspots, which control the smoothness of the signal. Such evolution, which reduces the sensitivity to long-period planet signatures, is no longer seen in the activity-induced Doppler-shifts, which is promising for long term RV monitoring surveys such as the Terra Hunting Experiment or the PLATO follow-up campaign.

**Key words:** line: profiles – methods: statistical – techniques: radial velocities – Sun: activity – planets and satellites: detection.

## 1 INTRODUCTION

Doppler spectroscopy is one of the bedrocks of past, present, and future exoplanetary science. As of 2024, it remains the preferred technique for confirming transiting planet candidates and the most prolific method for detecting non-transiting planets. It provides one of the most fundamental parameters of an exoplanet, its mass, crucial to constrain its inner composition (Mordasini et al. 2012) and to characterize its atmosphere (Batalha et al. 2019). Since the detection of 51 Pegasi b by Mayor & Queloz (1995), the radial velocity (RV) accuracy of optical high-resolution échelle spectrographs has dramatically increased, to the point where planets with RV semi-amplitudes below 1 m s<sup>−1</sup> can be reliably detected (e.g.

Faria et al. 2022; John et al. 2023). New-generation spectrographs like ESPRESSO (Pepe et al. 2021), EXPRES (Jurgenson et al. 2016), HARPS-3 (Thompson et al. 2016), NEID (Schwab et al. 2018), and KPF (Gibson et al. 2016) have been designed to provide RVs with sub 0.3-m s<sup>−1</sup> precision, suggesting that a detection of an Earth-like planet around a Sun analogue may be possible in the coming decade.

The intrinsic variability of stars is currently the main limitation to the detection of low-mass and long-period planets with Doppler spectroscopy (see Fischer et al. 2016; Crass et al. 2021; Meunier 2021). Various processes related to photospheric flows (e.g. granulation, supergranulation, meridional circulation; Dumusque et al. 2011; Meunier et al. 2015; Cegla et al. 2018; Meunier & Lagrange 2019, 2020) and magnetic activity (e.g. active regions, flares, magnetic cycles; Saar & Donahue 1997; Desort et al. 2007; Meunier, Desort & Lagrange 2010; Lovis et al. 2011; Gomes da Silva et al. 2012) distort stellar line profiles, giving rise to RV signals that hamper the search

\* E-mail: [baptiste.klein@physics.ox.ac.uk](mailto:baptiste.klein@physics.ox.ac.uk)

for planet signatures. The accurate modelling of these signals has become an extremely active area of research, with promising state-of-the-art methods currently under investigation (see Zhao et al. 2022, for a review). The mathematically tractable and flexible framework of Gaussian processes (Aigrain & Foreman-Mackey 2022) is now widely used to model the signals induced by stellar activity on the RVs and other activity indicators (e.g. Haywood et al. 2014; Rajpaul et al. 2015; Delisle et al. 2022; Barragán et al. 2022b). On the other hand, more and more innovative methods aim to correct line distortions directly from the cross-correlation functions (e.g. Collier Cameron et al. 2021; Klein et al. 2022; de Beurs et al. 2022) or from the spectra (e.g. Jones et al. 2017; Dumusque 2018; Rajpaul, Aigrain & Buchhave 2020; Cretignier, Dumusque & Pepe 2022; Lienhard et al. 2022). Yet, none of the methods is currently able to filter stellar activity RV signals significantly below the symbolic barrier of  $\sim 1 \text{ m s}^{-1}$  (after averaging over signals induced by P-mode oscillations and granulation), and remain poorly sensitive to low-amplitude ( $\lesssim 0.5 \text{ m s}^{-1}$ ) long-period ( $\gtrsim 100 \text{ d}$ ) planet signals.

The Sun is our best laboratory to better understand stellar activity signals in high-resolution spectra. It is the only star whose surface can be resolved at high resolution and for which planet-induced RV variations can be integrally removed. Since 2015, the Sun has been intensively monitored with the High Accuracy Radial-velocity Planet Searcher in the Northern hemisphere (HARPS-N; Cosentino et al. 2012), making it possible to investigate the full contributions of active regions on high-resolution spectra (Milbourne et al. 2019; Thompson et al. 2020; Haywood et al. 2022; Lienhard et al. 2023). These high-cadence high-precision solar spectra are ideally suited to assess the ability of a given method to filter stellar variability (Dumusque et al. 2021, Dumusque et al. submitted). Additionally, continuous monitoring of the solar surface with, for example, the Helioseismic and Magnetic Imager (HMI) instrument onboard the Solar Dynamics Observatory (SDO; Pesnell, Thompson & Chamberlin 2012) makes it possible to better understand the physical processes driving the observed disc-integrated RVs (e.g. Milbourne et al. 2019; Haywood et al. 2022, Rescigno et al., submitted).

In this study, we investigate how the HARPS-N solar activity RV signals can be modelled using information in both wavelength and time domains. In Section 2, we present our framework to separate a time series of cross-correlation functions (CCFs) into two components whose temporal variations are caused either by pure Doppler shifts or by shape-distortions. After describing the HARPS-N solar data set in Section 3, we investigate, in Section 4, the effects of filtering the shape-driven distortions on the solar RVs, and assess the planet detectability in the activity-filtered RVs through a large number of planet injection-recovery tests. In Section 5, we couple this framework with Gaussian processes and, notably, study the effect of combining newly extracted shape-driven activity indicators with the HARPS-N solar RVs in a multidimensional Gaussian process framework. Finally, in Section 6, we discuss the implications of our results in terms of activity modelling and planet search in the light of forthcoming long-term missions.

## 2 METHOD

Our approach builds on the Doppler-constrained principal component analysis (PCA) framework of Jones et al. (2017) and on the SCALPELS formalism introduced by Collier Cameron et al. (2021), borrowing elements from both to separate, as simply as possible, the components of the RV variations that can be explained by a pure Doppler shift from the rest. Using the same designations as Collier

Cameron et al. (2021), we refer to these two components as the shift- and shape-driven velocities, respectively.

### 2.1 Framework

We start from CCFs produced by, for example, the HARPS-N data reduction software (see Section 3.1). We consider a time series of  $n$  CCFs spanning  $m$  velocity bins  $\mathbf{v} = (v_1, \dots, v_m)$  and the associated RV time-series  $\mathbf{v}_{\text{obs}}$  (typically taken from the reduction pipeline). Our framework requires the computation of a non-noisy and activity-low reference CCF,  $\mathbf{C}_R$ , obtained by computing the inverse-variance-weighted mean of all the CCFs in the stellar rest frame.<sup>1</sup> We use the simple analytic framework of Jones et al. (2017) to decompose the time series of reference-subtracted CCFs,  $\mathbf{C}$ , into a pure Doppler component,  $\mathbf{C}_D$ , and residuals,  $\mathbf{C}_S$ , whose variations are driven by distortions in the profile shape. Assuming that the Doppler shifts are small compared to the line width, we approximate the Doppler component by the first-order term of the Taylor expansion of  $\mathbf{C}$  with respect to the CCF velocity bins,  $\mathbf{v}$  (see Bouchy, Pepe & Queloz 2001), matching to the spectrum's resolution. Therefore, noting  $\mathbf{D} = [\mathbf{C}_R'(v_1), \dots, \mathbf{C}_R'(v_m)]^T$ , the first derivative of the reference CCF with respect to the velocity bins, we have

$$\mathbf{C}_D = \mathbf{C} \cdot \frac{\mathbf{D}\mathbf{D}^T}{\mathbf{D}^T\mathbf{D}} \quad \mathbf{C}_S = \mathbf{C} - \mathbf{C}_D. \quad (1)$$

The time series of residuals,  $\mathbf{C}_S$ , contains all the shape distortions of the input CCFs, but are, in principle, insensitive to Doppler shifts. Following the method introduced by Collier Cameron et al. (2021), we use singular-value decomposition (SVD) to build a set of  $k$  independent vectors,  $\{\mathbf{U}_1, \dots, \mathbf{U}_k\}$ , from  $\mathbf{C}_S$ . Each of these vectors is a time-series of  $n$  points associated with a score that scales with its contribution to the variance of  $\mathbf{C}_S$ . In what follows, the basis vectors are sorted by decreasing score value, so that  $\mathbf{U}_1$  and  $\mathbf{U}_k$  contribute the most and the least to the variance of  $\mathbf{C}_S$ , respectively.

Finally, we project our reference-subtracted CCFs  $\mathbf{C}$  onto the basis  $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_k)$ , where each column  $i$  is equal to vector  $\mathbf{U}_i$ . Using the same notations as Collier Cameron et al. (2021), we call  $\mathbf{C}_{\parallel}$  the projection of  $\mathbf{C}$  on  $\mathbf{U}$ , such that

$$\mathbf{C}_{\parallel} = \mathbf{U} \cdot \mathbf{U}^T \cdot \mathbf{C}. \quad (2)$$

We extract a RV time-series,  $\mathbf{v}_{\parallel}$  (called shape-driven RVs), by fitting a Gaussian function to each CCF of  $\mathbf{C}_{\parallel} + \mathbf{C}_R$ . Finally, we build a Doppler-driven (hereafter, called shift-driven) RV time-series,  $\mathbf{v}_{\perp}$ , by subtracting  $\mathbf{v}_{\parallel}$  from the initial RV time-series  $\mathbf{v}_{\text{obs}}$ .

The number  $k$  of principal components used to build  $\mathbf{U}$  must be chosen with caution. Basis vectors marginally contributing to the variance of the input CCF time-series are primarily composed of white noise. Thus, including them in the projection (equation 2) will increase the dispersion of the input RVs and interfere with potential planetary signals. In practice, we adapted the method described in the section 3 of Klein et al. (2024) to choose the number of components used in the projection. We independently generate about 10 matrices, matching  $\mathbf{C}_S$  in shape, but containing only white noise, drawn from the formal uncertainties on the CCF. We then apply SVD to each of these matrices and use the highest eigenvalue  $S_{\text{max}}$  of all matrices as a

<sup>1</sup>One alternative could be to compute  $\mathbf{C}_R$  by averaging just a subset of CCFs, around the activity minimum of the star. However, we caution that robustly identifying activity minimums can be challenging in some cases, and that stars still exhibit signs of activity during the cycle minimum, as shown in Section 5.2. In practice, we find that the computation of  $\mathbf{C}_R$  marginally impacts the results.

**Table 1.** Description of the parameters used to generate time series of CCFs with SOAP2. In each case, we give the number of spots ( $N_{\text{spot}}$ ) and faculae ( $N_{\text{fac}}$ ), as well as their corresponding stellar rotational phase  $\phi_r$  and latitudes (Lat). The size of the active regions are set to 0.18 % (faculae) and 0.1 % (spot) with respect to the area of the visible hemisphere, in order to produce an RV signal of about  $1 \text{ m s}^{-1}$ . The last column gives the number  $N_{\text{pl}}$  of pure Doppler signals injected in the CCFs.

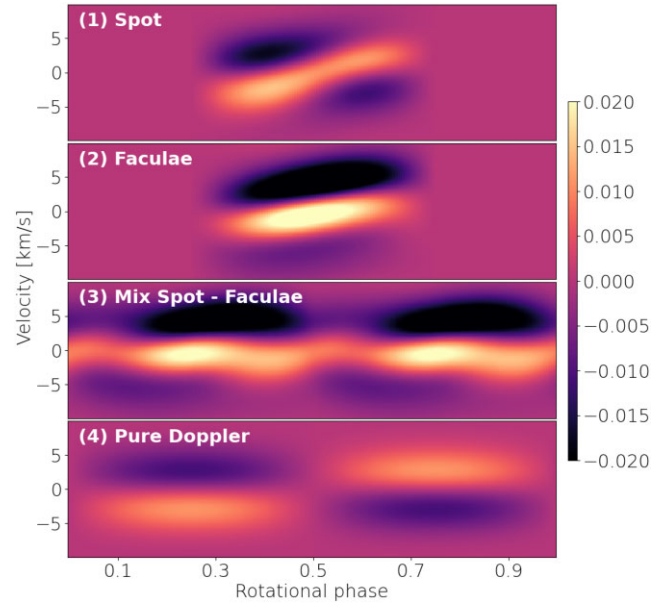
Case	$N_{\text{spot}}$	$\phi_r$	Lat	$N_{\text{fac}}$	$\phi_r$	Lat	$N_{\text{pl}}$
–	–	–	[deg]	–	–	[deg]	–
(1)	1	0.5	0	0	–	–	0
(2)	0	–	–	1	0.5	0	0
(3)	2	0.0, 0.5	30	2	0.25, 0.75	30	0
(4)	0	–	–	0	–	–	1

proxy for noise-dominated components. When we apply SVD to  $\mathbf{C}_S$ , components with eigenvalues greater than  $S_{\text{max}}$  are assumed to enclose a significant fraction of correlated noise (e.g. activity-induced shape distortions) and are used in the basis definition. Conversely, components with eigenvalues smaller than  $S_{\text{max}}$  are assumed to be dominated by white noise and discarded. Note that, in practice, the selected principal components used in the basis are consistent with the ‘elbow’ heuristic method and account for an explained variance score typically greater than 95 per cent. Assuming that  $\mathbf{U}$  is noise free, all RV variations induced by distortions of the input CCFs are enclosed in  $\mathbf{v}_{\parallel}$ . Therefore, in principle,  $\mathbf{v}_{\perp}$  will be composed of planet signatures, Doppler shifts of unknown origin (e.g. residual stellar activity or instrument systematics) and white noise. As a word of caution, we stress that our theoretical framework is limited to spots and faculae, which do affect the shape of the CCFs. Other processes like supergranulation or meridional circulation may induce Doppler shifts without distorting the CCFs (see Meunier 2021, and references therein).

## 2.2 Application to synthetic data

Before applying the framework introduced in Section 2.1 to real solar data, we conduct some simple tests on synthetic CCFs computed using the SOAP2 online tool (Boisse, Bonfils & Santos 2012; Dumusque, Boisse & Santos 2014). This code allows the user to simulate the effects of spots (i.e. pure brightness inhomogeneities) and faculae (via the inhibition of the convective blueshift) on the absorption line profile. We consider three different configurations of active regions, listed in Table 1. Cases (1) and (2) correspond to a single spot and a single facula, at the equator of the star. For case (3), we consider a mix of four active regions, two spots and two faculae, of equal latitude ( $30^\circ$ ) and evenly spread in stellar longitude. Following Berdyugina (2005) and Meunier et al. (2010), the temperature of the spots is set to  $\sim 650 \text{ K}$  below the Sun’s effective temperature. The size of the spots (resp. faculae) is set to 1 per cent (resp. 0.18 per cent) of the visible stellar hemisphere, so that the semi-amplitude of the corresponding RV signal is about  $1 \text{ m s}^{-1}$ , which is the typical order of magnitude of activity-induced RV variations for the Sun (e.g. Meunier et al. 2010).

Finally, in case (4), we consider the case of a pure Doppler signal in absence of stellar activity, in order to double check that our method does not affect planetary signatures. The CCFs are generated by interpolating a non-spotted line profile computed with SOAP2, and shifting the entire profile according to the RV signature induced by a  $1 \text{ m s}^{-1}$  sine-wave with a period equal to the stellar rotation period to the data (note that, no stellar activity is considered in this case). In all cases, we generate a time-series of 200 CCFs

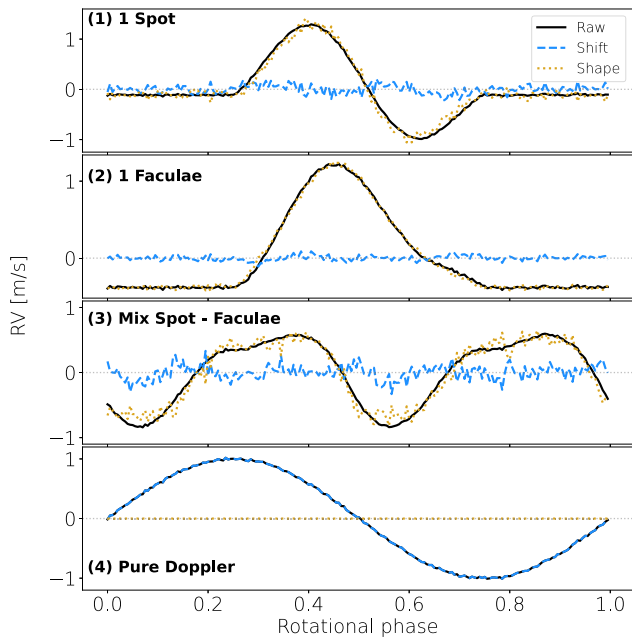


**Figure 1.** Dynamical spectrum of the different time series of CCFs generated with SOAP2 using the parameters listed in Table 1. Each panel depicts the relative flux of the synthetic CCFs (after subtraction of a reference line profile, computed in absence of activity), in per cent, as a function of the rotational phase of the star ( $x$ -axis) and line velocity ( $y$ -axis).

evenly spanning one stellar rotation cycle, and add a white noise of  $10^{-5}$  with respect to the normalized continuum, which is of the same order of magnitude as the photon noise in HARPS-N solar CCFs. Note that, adding white noise significantly larger than the planet-induced Doppler shift is crucial in case (4). Since the Taylor expansion of equation (1) is only an approximation of the first derivative of the CCFs with respect to the wavelength, a small fraction of the planet signature will still be present in the residuals of Taylor expansion (i.e. the matrix  $\mathbf{C}_S$  in Section 2.1). If the amplitude of this contribution is significantly weaker than the noise, it will naturally be discarded during the dimension reduction process (i.e. the SVD). For example, a  $1 \text{ m s}^{-1}$  Doppler signal induces residuals of about  $10^{-8}$  in  $\mathbf{C}_S$ , about 3-to-4 orders of magnitude lower than the typical white noise in HARPS-N spectra. A way to account for these variations, described in Collier Cameron et al. (2021), Wilson et al. (2022), and John, Collier Cameron & Wilson (2022), is to perform the dimensional reduction and search for planet-induced Doppler signatures simultaneously, which we apply in Section 4.2. Note however that, the planet signatures considered in this study are small enough not to affect significantly the shape-driven components of the Taylor expansion.

We then apply the framework of Section 2.1 to the simulated data. We use the unspotted line profile generated by SOAP2 as reference ( $\mathbf{C}_R$  in Section 2.1). A total of 3, 4, and 4 principal components is used to construct the basis  $\mathbf{U}$  in case (1), (2), and (3), respectively. As for case (4), eigenvectors are found to be dominated by white noise and are therefore discarded, which is expected as no shape distortion was introduced in the data set. In this last case, adding components in the matrix  $\mathbf{U}$  introduces white noise in both  $\mathbf{v}_{\parallel}$  and  $\mathbf{v}_{\perp}$ , but the amplitude of the planet signature is only affected starting from  $\mathbf{U}_{20}$  ( $\sim 3$  per cent of the explained variance in a data set dominated by white noise). The time-series of noise-free reference-subtracted CCFs are shown in Fig. 1 for the cases listed in Table 1. The shape- and shift-driven RVs (i.e. resp.  $\mathbf{v}_{\parallel}$  and  $\mathbf{v}_{\perp}$  in Section





**Figure 2.** Time-series of raw RVs (black solid lines), shift-driven RVs ( $v_{\perp}$ ; blue dashed lines) and shape-driven RVs ( $v_{\parallel}$ ; yellow dotted lines), as a function of the rotational phase of the star, for the simulations described in Table 1.

2.1) extracted in each case are shown in Fig. 2. Unsurprisingly, the Doppler component is well recovered in case (4). We also find that RV signals induced by active regions are well filtered, with a suppression of 86, 94, and 78 per cent of the stellar activity signals in cases (1), (2), and (3), respectively. However, even in the ideal case considered in this section,  $v_{\perp}$  still exhibits activity-induced fluctuations of up to  $0.1 \text{ m s}^{-1}$  RMS in case (3), significantly larger than the level of white noise ( $\sim 0.01 \text{ m s}^{-1}$  RMS). This is due to the fact that Doppler shifts and shape distortions are not purely orthogonal in the Taylor expansion and therefore cannot be entirely separated in equation (1). This can be understood by the fact that the first derivative of the CCF has no reason to be orthogonal to higher-order derivatives (especially with odd-order derivatives). By analysing the periodicities in  $v_{\parallel}$  and  $v_{\perp}$  for cases (1) and (2), we find that, in the case of a single spot, the shift-driven RVs,  $v_{\perp}$ , still exhibits modulations at harmonics of the star’s rotation period, whereas, in the case of a single facula,  $v_{\perp}$  does not exhibit any explicit sign of periodicity. Finally, we note that including more SVD components in the basis projection affects only marginally the recovered RV signals and that none of the additional principal components exhibit significant periodic modulation.

### 3 DATA

#### 3.1 Solar observations and data selection

Since 2015, disc-integrated solar spectra were collected at a 5-min cadence with the stabilized cross-dispersed échelle spectrograph HARPS-N (Cosentino et al. 2012; Dumusque et al. 2015; Phillips et al. 2016; Collier Cameron et al. 2019; Dumusque et al. 2021), mounted at the 3.58-m Telescopio Nazionale Galileo (TNG) at Roque de Los Muchachos observatory (La Palma, Spain). These observations leverage the high-resolving power ( $\mathcal{R} = 115\,000$ ) and the large spectral coverage in the V band (from 383 to 690 nm) of

the instrument to achieve RV uncertainties lower than  $0.5 \text{ m s}^{-1}$ . Our input data set was collected between 2015 July and 2023 September and processed with the ESPRESSO data-reduction software (DRS v.2.3.5; Pepe et al. 2021), adapted to the HARPS-N solar telescope in Dumusque et al. (2021) and Dumusque et al., submitted.

Since the Sun’s surface is resolved, any partial obstruction of the solar disc by, for example, clouds in Earth’s atmosphere, will likely induce substantial distortions in the CCFs, and, thus, large spurious RV deviations. In order to flag cloud-contaminated spectra, Collier Cameron et al. (2019) perform a daily linear fit to the apparent solar magnitude as a function of airmass (corresponding to the expected extinction law in optical observation conditions). The magnitude is given, as in Collier Cameron et al. (2019), by  $m = -5 \log_{10} S/N_{60}$ , where  $S/N_{60}$  is the signal-to-noise ratio (S/N) in the 60th spectral order (corresponding to echelle order 98 which has a central wavelength of 625 nm). The deviation of each data point to the best-fitting extinction law is estimated through a Bayesian mixture model (Hogg, Bovy & Lang 2010), assigning a probability  $p$  that the spectrum is not contaminated by clouds. Following the recommendations of Dumusque et al. (2021) and Collier Cameron et al. (2021), we only consider spectra with  $p > 0.95$  in the following analysis and airmass observations lower than 2.25. After rejecting such spectra, our initial data set of 147 741 solar spectra is reduced down to 104 573 (70.8 per cent). The median S/N in the 60th spectral order is 355 and the median RV uncertainty per spectrum is  $0.28 \text{ m s}^{-1}$ . The fraction of retained spectra are in agreement with Al Moulla et al. (2023), where in their fig. A.2 it is shown that for most days there are only about 0–2 outliers per day among the dozens of daily HARPS-N solar spectra. By keeping only the spectra least affected by clouds, we are able to compute daily stacked spectra (see the following section) that are minimally affected by differential extinction.

#### 3.2 Spectra post-processing

The analysis of solar activity with HARPS-N is made complex since Sun-as-a-star observations are different from real stellar observations. For instance, since the Earth is turning around the Sun and since the Earth is not perfectly aligned with the solar equator, a substantial change in the Sun’s projected rotational velocity ( $v \sin i$ ) is observed, which introduces 1-year and half-a-year signals in the time-series of the full-width at half maximum (FWHM) of the CCFs (Collier Cameron et al. 2019). Furthermore, the HARPS-N cryostat was changed on the 2021 October 4th, which also introduced an offset in several CCF moments since the focus of the instrument was realigned, slightly improving the resolution. Those instrumental interventions are expected to have RV contributions smaller than  $0.5 \text{ m s}^{-1}$  since no clear offset is observed by eye in either the solar RVs or standard stars RVs, but likely affected the spectra in some way.

In order to remove those signals and better isolate stellar activity contributions, we post-processed the SID spectra time-series with YARARA (Cretignier et al. 2021). Note that, we did not use any time-domain empirical corrections developed for the atmospheric extinction or  $v \sin i$  (Collier Cameron et al. 2019; Dumusque et al. 2021), since those corrections were developed for the time-domain and their ability to perform in the wavelength domain is unclear. However, as shown below, YARARA itself is able to correct for them. YARARA is a post-processing methodology that aims to split the different contaminations coming from the instruments, such as ghosts, stitchings, interference patterns, ThAr bleeding, point spread

function (PSF) defocus, from the tellurics and from the stellar activity directly in the spectra.

The main steps of the pipeline are to (i) daily stack the SID spectra in the heliocentric rest-frame, (ii) normalize the continuum of the spectra with RASSINE (Cretignier et al. 2020b), and (iii) correct the continuum-normalized spectra with YARARA (Cretignier et al. 2021, 2023). The data set after daily stacking the spectra and removing anomalous residuals among them.<sup>2</sup> consists of 1880 daily stacked spectra. Note that, we only use the first version of YARARA (hereafter ‘YV1’), which performs systematics correction in the wavelength domain as described in Cretignier et al. (2021). The pipeline was slightly upgrade on HARPS-N to better correct the cryostat change (see Appendix A).

The final RVs were obtained from CCFs with a line list tailored for the Sun as described in Cretignier et al. (2020a), formally described as ‘Custom’ in Bourrier et al. (2021) and ‘Kit-Cat’ line list in Cretignier (2022). The CCF’s FWHM, bisector velocity span, ( $V_s$ ; Hatzes 1996; Queloz et al. 2001), and equivalent width (EW; product of FWHM and contrast), as well as the Mount Wilson S index ( $S_{HK}$ ; Noyes et al. 1984) are also computed from the processed spectra. The CCF EW (i.e. the area of the CCF) is a known tracer of stellar temperature and metallicity (Malavolta et al. 2017). As the Sun’s metallicity is expected to remain roughly constant during our observations, the CCF EW is a good tracer of global temperature changes in the solar photosphere (e.g. magnetic intensification and/or spot-induced flux variations; Collier Cameron et al. 2019). On the other hand,  $V_s$ , defined as the RV difference between the centre of a Gaussian fit to the whole CCF and the centre of a parabola fit on the core of the CCF (i.e. within  $\pm 2.3 \text{ km s}^{-1}$ ), is highly sensitive to velocity suppressions on the solar disc (mostly due to faculae).

Because YARARA is working with spectra interpolated on a 0.01-Å wavelength grid, the CCFs are slightly oversampled to velocity bins of  $530 \text{ m s}^{-1}$ , compared to the  $820 \text{ m s}^{-1}$  of the usual DRS (Dumusque et al. 2021). This is not critical, but it means that two consecutive velocity bins are not fully independent and share some covariance. Flux uncertainties on the CCF are artificially boosted by the square-root of the oversampling factor to cancel the oversampling effect. The natural output of the YARARA post-processing (spectra, CCFs or time-series) is usually already corrected from the stellar activity component. Since the purpose of this work is to study the solar activity signals, we introduced back the correction related to activity by YARARA to form an ‘YVA’ data set in order to follow the terminology introduced in Dalal, Rescigno & Cretignier (in preparation). The time-series obtained with the YVA data set are shown in Fig. 3.

## 4 RESULTS

### 4.1 Planet-free case

We apply the framework defined in Section 2.1 to the time-series of daily binned solar CCFs. Using the mean line profile as reference profile  $C_R$ , we compute the time series of reference-subtracted CCFs (see the first panel of Fig. 4). From equation (1), we derive a time-series of shift-driven (i.e. Doppler component in Section 2.1) and shape-driven CCFs, shown in the middle and bottom panels of Fig. 4. As expected, the Doppler component only contains information on the first derivative of the CCFs, as evidenced by the fact that, at

each epoch, the CCF variations are symmetrical with respect to the centre of the line (i.e. zero-velocity point). On the other hand, shape-driven residuals exhibit a more complex structure, non symmetrical with respect to the line centre, suggesting that it probes higher-order derivatives of the CCF variations. We also double checked that the width and asymmetry of the shift-driven profiles remain constant over time.

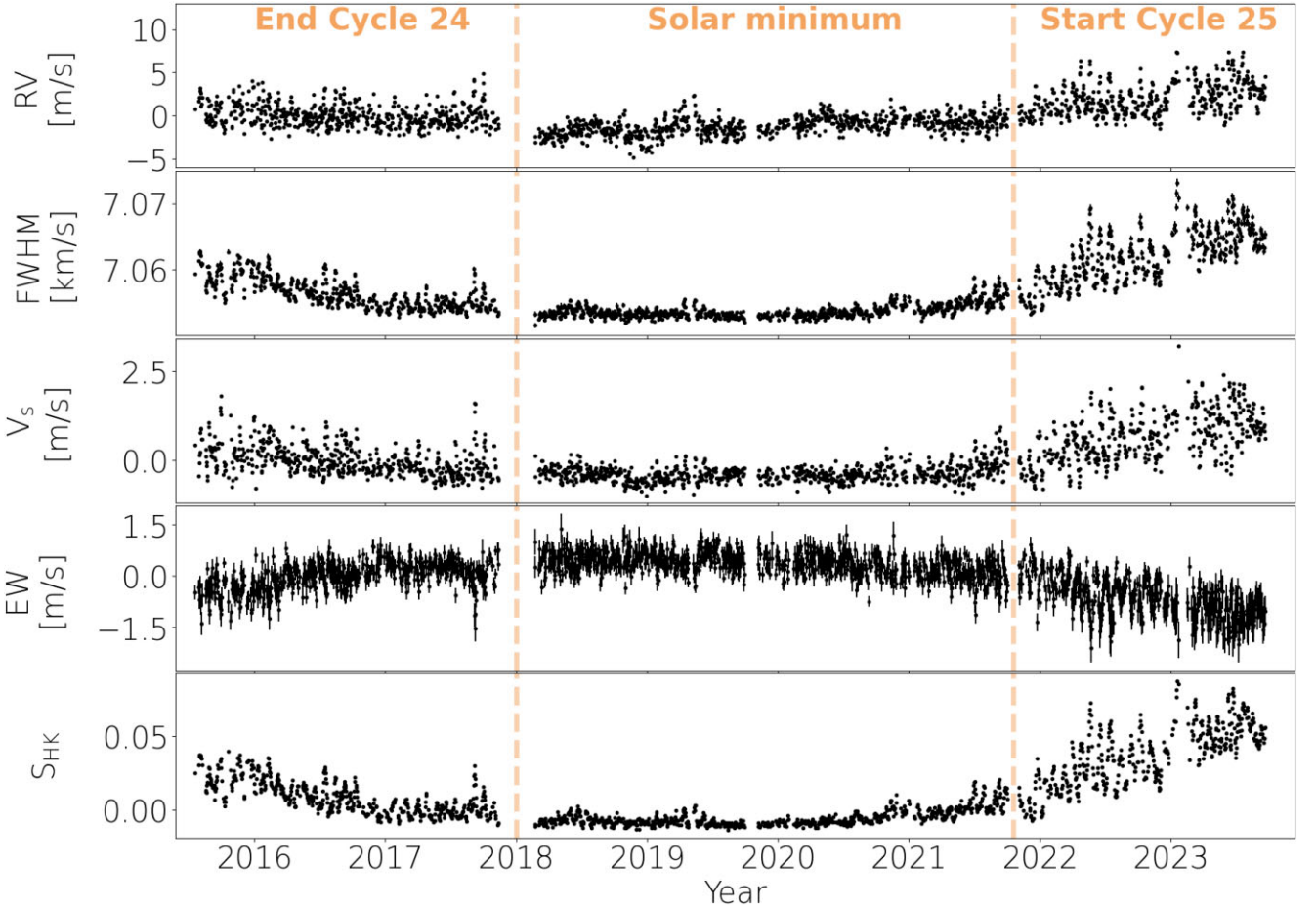
We then apply SVD to the shape-driven CCFs and extract a basis of orthogonal vectors tracing the main non-Doppler-induced line profile variations. Applying the procedure described in Section 2, we use the first four principal components in our reconstruction. As shown in Fig. 5, the first eigenvector  $U_1$  correlates strongly with  $v_{\text{obs}}$  and usual activity indicators. In particular, we obtain a weighted Pearson correlation coefficient of 0.90 between  $U_1$  and  $S_{HK}$ , which suggests that the variations of these two quantities are driven by similar processes (see Section 5 for a more detailed comparison). Higher-order SVD components tend to exhibit weaker correlations with CCF indicators, with the exception of  $U_4$ , which exhibits a relatively strong anticorrelation with the EW of the CCF, indicating that this indicator could be a good proxy of the photospheric temperature (see Collier Cameron et al. 2019).

We project the input CCFs onto the basis defined by components  $U_1$  to  $U_4$ , and derived shape- and shift-driven RVs (resp.  $v_{\parallel}$  and  $v_{\perp}$ ) using the method described in Section 2.1. Both time-series are shown in Fig. 6 and their root-mean-square (RMS) deviations are given in Table 2. Starting from a dispersion of  $2.1 \text{ m s}^{-1}$  in the input data, the shape- and shift-driven RVs exhibit RMS dispersions of  $1.8$  and  $1.1 \text{ m s}^{-1}$ , respectively.<sup>3</sup> In order to assess the effectiveness of the method for different levels of solar activity, we divide our input data set in the three seasons shown in Fig. 3. From 2015 to 2018 (Season 1), the Sun reaches the end of its activity cycle but still exhibits clear signs of activity. The Sun enters then a long period of activity minimum (Season 2, 2018–2021.8) which lasts until the end of 2021. With the start of Cycle 25 (Season 3, 2021.8–2024), the Sun exhibits activity signals with significantly larger amplitude than in Season 1. Note that, this division is motivated by two factors. First, the cycle-induced evolution of the solar activity, evidenced, for example, by the amplitude of the fluctuations of the disc-integrated magnetic flux density (lower than  $0.5 \text{ G}$  in Season 2, according to SDO/HMI data extracted with SOLASTER; see Ervin et al. 2022) allows to separate three different activity regimes. Secondly, important instrument maintenance operations, namely the change of Fabry–Pérot interferometer, between Seasons 1 and 2, and the refurbishment of the CCD camera, between Seasons 2 and 3, justify the definition of the three seasons adopted in this study. From Table 2, we note that our activity-filtering framework performs best when the Sun is more active, with a 44 per cent reduction in RV RMS in Season 3, compared to only 7 per cent in Season 2 (solar minimum). We also note that, in Season 1, our results are similar to those obtained with SCALPELS, on a similar data set (Collier Cameron et al. 2021).

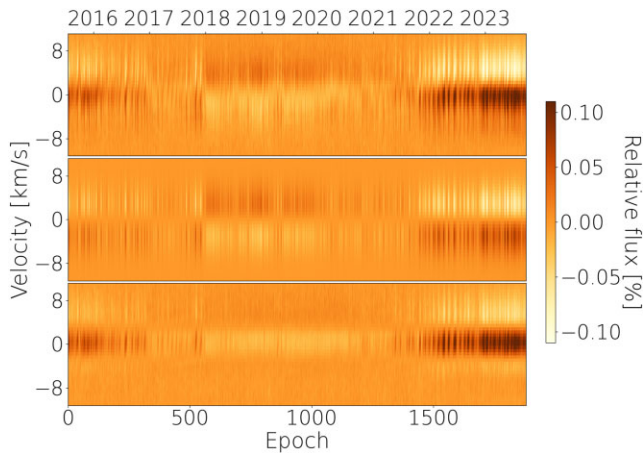
As shown in Fig. 6, the shape-driven RVs seem to enclose most of the long-term RV variations, primarily driven by the magnetic cycle. This is further evidenced in the generalized Lomb–Scargle (GLS; Zechmeister & Kürster 2009) periodograms of the RV time-series (see bottom curve on the top panel of Fig. 7), which are dominated by peaks at periods larger than  $\sim 100 \text{ d}$ . We note also a significant power near the Earth orbital period, also present in the periodogram

<sup>2</sup>Anomalous residuals spectra can be due to spectra with unperfect corrections or lower SNR observations.

<sup>3</sup>Note that, the quadratic sum of the RMSs of the shift- and shape-driven RVs is not strictly equal to the RMS of  $v_{\text{obs}}$ , which reflects the fact that the decomposition defined in Section 2.1 is not entirely orthogonal.



**Figure 3.** YVA solar data set. From top to bottom: time-series of HARPS-N solar RVs, FWHM, velocity span and equivalent width of the HARPS-N solar CCFs, and S index. The vertical dashed lines delimitate the three seasons defined in Table 2.



**Figure 4.** Time-series of reference-subtracted Solar CCFs (top panel), Doppler components computed using equation (1) ( $C_D$ , middle panel), and shape-driven residuals ( $C_S$ , bottom panel).

of component  $U_2$ , attributable to residuals of the effects of Earth’s orbital eccentricity and solar obliquity on the FWHM of the CCF, as described in Collier Cameron et al. (2019). These effects will not be observed in other stars. We also note that the power at the Sun’s rotation period and harmonics remains strong in the shift-

driven RVs, despite a fraction of it being transferred to the shape-driven RVs. Surprisingly, components  $U_3$  and  $U_4$  exhibit a significant modulation of the Sun’s rotation period and first harmonics and, thus, could potentially act as good proxies of quasi-periodic stellar activity signals (see Section 5), although no strong correlation with CCF activity indicators is observed.

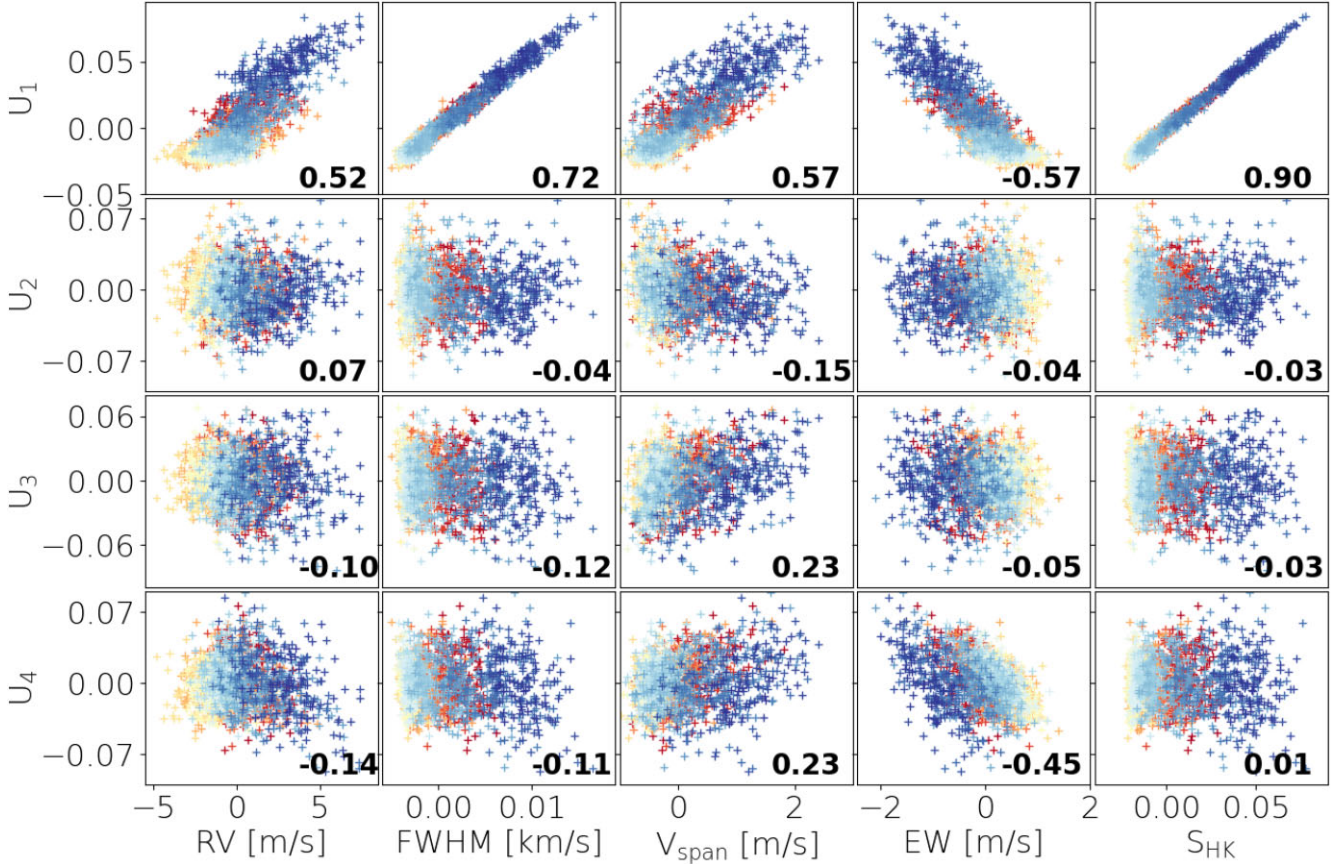
#### 4.2 Effects on planet recovery

In order to assess the ability of our activity-filtering procedure to preserve planet RV signatures, we create 1000 data sets from the solar HARPS-N observations, each containing a single planetary signal directly injected in the CCF time-series with the same temporal sampling. As we assume the planet orbit to be circular, the corresponding RV signature,  $v_p$ , as a function of the time,  $t$ , is given by

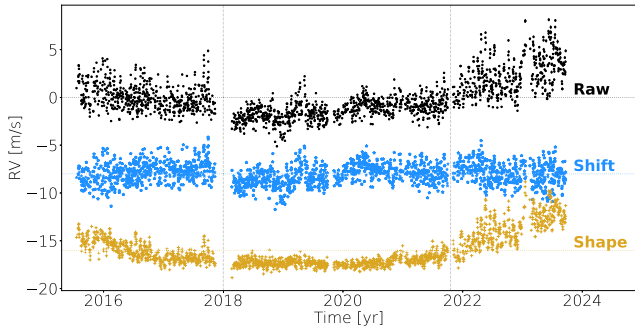
$$v_p(t) = K_{\text{inj}} \sin 2\pi \left[ \frac{T_0 - t}{P_{\text{orb}}} + \phi_p \right] \quad (3)$$

where the reference time,  $T_0$ , is set to  $\text{BJD} = 2457\,223.490\,54$  (i.e. the time of our first observation), and where  $K_{\text{inj}}$ ,  $P_{\text{orb}}$  and  $\phi_p$  are respectively the semi-amplitude, orbital period and orbital phase of the injected signature. These three parameters are randomly drawn using log-uniform laws (for  $K_{\text{inj}}$  and  $P_{\text{orb}}$ ) and uniform laws (for  $\phi_p$ ), as described in Table 3.





**Figure 5.** First four SVD components ( $U_1$  to  $U_4$ ) plot against the median-subtracted HARPS-N solar RVs, the FWHM, bisector velocity span and EW of the CCFs, and the  $S_{HK}$  time series. The individual points are colour-coded by date of observation, bluer (redder) points corresponding to more recent (older) observations. In the bottom right of each panel, we indicate the Pearson correlation coefficient of the two time-series (weighed by the uncertainties on the x-axis). Note that, all correlation coefficients are associated with  $p$ -values lower than 0.05, and are thus considered to be significant.



**Figure 6.** Time series of HARPS-N solar RVs (top curve), shift-driven RVs ( $v_{\perp}$ ; curve in the middle), and shape-driven RVs ( $v_{\parallel}$ ; bottom curve). The typical formal error bar on each RV point is about  $0.16 \text{ m s}^{-1}$ . The vertical dashed lines divide the data into the three periods shown in Fig. 3.

To inject the planet signature, we interpolate each CCF in the solar rest frame using a centered square-exponential Gaussian Process (GP; Rasmussen & Williams 2006; Aigrain & Foreman-Mackey 2022) with covariance function  $k$  between each pair of velocity bins  $v_i$  and  $v_j$  given by

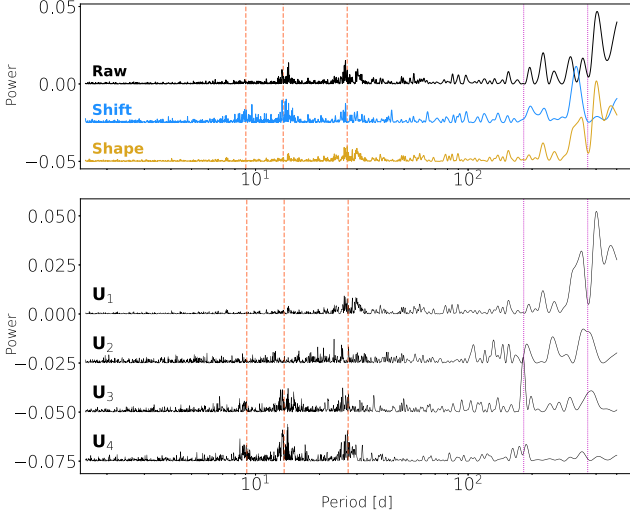
$$k(v_i, v_j) = A \exp \left[ -\frac{(v_i - v_j)^2}{2\lambda_c^2} \right] + \eta_i^2 \delta_{i,j} \quad (4)$$

**Table 2.** RMS of the time series of  $v_{\text{obs}}$ ,  $v_{\perp}$  (shift-driven RVs) and  $v_{\parallel}$  (shape-driven RVs), over different periods of time. In each case, the first four principal components are used to build the time-series of shape-driven CCFs,  $C_{\parallel}$ , from which  $v_{\parallel}$  is extracted.

Name	Period	$v_{\text{obs}}$	$v_{\perp}$	$v_{\parallel}$
Season 1	2015–2018	$1.34 \text{ m s}^{-1}$	$1.08 \text{ m s}^{-1}$	$0.97 \text{ m s}^{-1}$
Season 2	2018–2021.8	$1.07 \text{ m s}^{-1}$	$1.00 \text{ m s}^{-1}$	$0.50 \text{ m s}^{-1}$
Season 3	2021.8–2023.7	$1.98 \text{ m s}^{-1}$	$1.11 \text{ m s}^{-1}$	$1.69 \text{ m s}^{-1}$
Altogether	2015–2023.7	$2.05 \text{ m s}^{-1}$	$1.06 \text{ m s}^{-1}$	$1.76 \text{ m s}^{-1}$

where  $\eta_i$ , is the uncertainty on the CCF flux associated with pixel  $i$ , and  $\delta$  stands for the Kronecker delta. The two free hyperparameters of equation (4), namely the GP amplitude  $A$  and correlation length  $\lambda_c$ , are estimated by maximizing the likelihood of the data, computed using the GEORGE PYTHON module (Ambikasaran et al. 2015). As a sanity check, we measured the RV and FWHM of each planet-injected CCF by fitting a Gaussian profile to it. The resulting RV/FWHM time-series differ by no more than  $0.1\sigma$  from their DRS-provided counterparts, confirming that this interpolation procedure only marginally affects the shape and position of the line profile. As a word of caution, we note that injecting the planet signature at the CCF level rather than at the spectrum level assumes that the Doppler shift is integrally preserved in the cross-correlation process. This assumption is motivated by recent studies demonstrating that





**Figure 7.** Top panel: GLS periodogram of the HARPS-N solar RVs (top line), the shift-driven RVs ( $v_{\perp}$ , middle line) and shape-driven RVs ( $v_{\parallel}$ , bottom line). Bottom panels: GLS periodogram of the first four principal components,  $U_1$  to  $U_4$ , of the shape-driven CCFs. From right to left, the vertical orange dashed and magenta dotted lines indicate the Solar rotation period (and first two harmonics) and the Earth orbital period (and first harmonic). The periodograms were computed using the ASTROPY PYTHON module (Astropy Collaboration et al. 2013, 2018, 2022).

**Table 3.** Probability laws used to generate the semi-amplitude, orbital period and orbital phase of the planet RV signatures injected in the HARPS-N Solar CCFs (see Section 4.2).  $\mathcal{U}$  and  $\log \mathcal{U}$  stand respectively for the Uniform and log-Uniform probability laws. The last two columns give the minimum and maximum values used for each parameter.

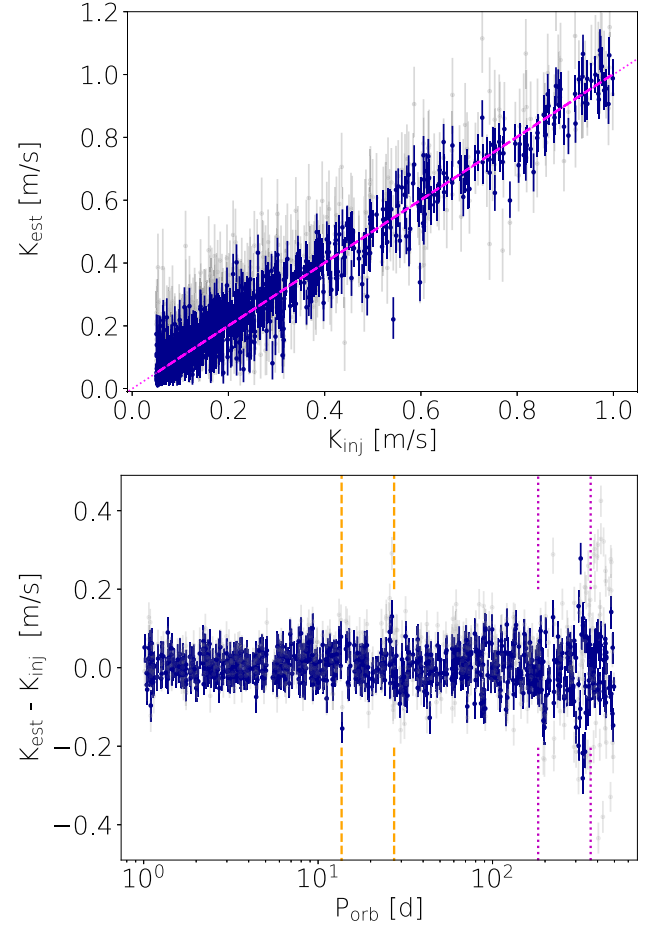
Parameter	Notation	Probability law	Min	Max
Semi-amplitude	$K_{\text{inj}}$	$\log \mathcal{U}$	$0.05 \text{ m s}^{-1}$	$5.0 \text{ m s}^{-1}$
Orbital period	$P_{\text{orb}}$	$\log \mathcal{U}$	1.0 d	500 d
Orbital phase	$\phi_p$	$\mathcal{U}$	0.0	1.0

reliable planet parameters were extracted from the CCF (e.g. John et al. 2023; de Beurs et al. 2024).

Our procedure to retrieve the injected planet signatures is similar of that used in Collier Cameron et al. (2021) and Wilson et al. (2022). For each data set, we jointly apply the dimensional reduction of Section 2.1 and fit for a planet Doppler motion with equation (3). As detailed in the appendix A of Wilson et al. (2022), projecting the observed RVs  $v_{\text{obs}}$  onto the matrix  $\mathbf{P} = \mathbf{I} - \mathbf{U} \cdot \mathbf{U}^T$ , where  $\mathbf{I}$  is the identity matrix and  $\mathbf{U}$  the basis defined in Section 2.1 (using four principal components), is equivalent to subtracting  $v_{\parallel}$  from  $v_{\text{obs}}$ . The likelihood  $\mathcal{L}$  of the data given the model parameters  $\theta$  is then given by

$$\ln \mathcal{L} = -\frac{1}{2} [\mathbf{P} \cdot (v_{\text{obs}} - v_p(\theta))]^T \cdot \Sigma^{-1} [\mathbf{P} \cdot (v_{\text{obs}} - v_p(\theta))] - \frac{1}{2} \ln |\Sigma| - \frac{N}{2} \ln 2\pi \quad (5)$$

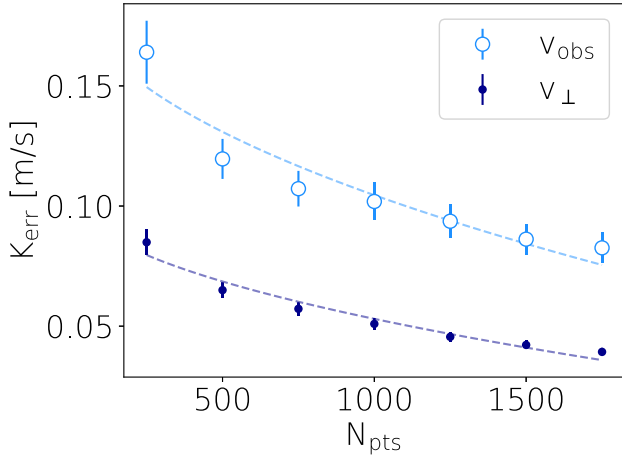
where  $v_p$  is the planet RV model given by equation (3) and  $N$  is the number of data points. The covariance matrix  $\Sigma$  is assumed diagonal, with  $\Sigma_{kk} = \sigma_k^2 + \sigma_j^2$ , where  $\sigma_k$  is the formal RV uncertainty on the  $k$ th data point and  $\sigma_j$  is a free parameter of the model to absorb RV variations that are not captured by equation (3) or the formal RV uncertainties (e.g. activity and granulation signals, instrumental stability). We assume that the planet orbital period and phase are known, leaving the planet RV semi-



**Figure 8.** Top: Planet RV semi-amplitude,  $k_{\text{est}}$ , recovered from the activity-filtered RVs ( $v_{\perp}$ , dark blue dots) and from the raw RVs ( $v_{\text{obs}}$ , light grey dots), as a function of the RV semi-amplitude  $K_{\text{inj}}$  of the planets injected in Section 4.2. The magenta dotted and dashed lines indicate the identity and the best-fitting straight line, respectively. Bottom: Difference between  $k_{\text{est}}$  and  $K_{\text{inj}}$  as a function of the orbital period  $P_{\text{orb}}$  of the injected planet. The vertical orange dashed and magenta dotted lines indicate the Solar rotation period (and first harmonic) and the Earth orbital period (and first harmonic).

amplitude and the jitter term  $\sigma_j$  as the only free parameters of the model. Their posterior probability given the data is estimated, in the Bayesian framework, using the affine-invariant Markov Chain Monte Carlo (MCMC) process EMCEE (5000 iterations of 100 chains; Foreman-Mackey et al. 2013). The best-fitting parameters and  $1\sigma$ -uncertainties are estimated from the median and from the 16th and 84th percentiles of the chain after removing a burn-in period significantly longer than the autocorrelation time of the chain (about 100 iterations). As a reference, we also fit equation (3) to the input RV time-series,  $v_{\text{obs}}$ , before filtering the shape-induced contributions.

The distribution of recovered RV semi-amplitudes,  $k_{\text{est}}$ , is shown as a function of  $K_{\text{inj}}$  in the top panel of Fig. 8 for  $K_{\text{inj}} < 1 \text{ m s}^{-1}$ . In most cases, the recovered values of the planet RV semi-amplitude match their injected counterpart, with a slope of  $1.03 \pm 0.1$ . As shown in the bottom panel of Fig. 8, deviations from the identity are observed when the planet orbital period lies close to the Sun's rotation period or to the Earth orbital period (and first harmonic), which roughly corresponds to the location of the peaks in the periodogram of  $v_{\perp}$  (see Fig. 7). No particular trend is observed



**Figure 9.** Mean error  $K_{\text{err}}$  on the planet RV semi-amplitude estimated from the activity-filtered RVs ( $v_{\perp}$ , dark blue dots) and from the raw RVs ( $v_{\text{obs}}$ , light blue open circles) as a function of the number of data points  $N_{\text{pts}}$ . For each case, the dashed line indicates the best-fitting square-root law (reduced  $\chi^2$  of 1.0 and 1.1 for  $v_{\text{obs}}$  and  $v_{\perp}$ , respectively).

between  $k_{\text{est}}$  and the injected planet RV semi-amplitude or orbital phase.

Let us assume, conservatively, that a planet is detected if (i)  $k_{\text{est}}$  differs from  $K_{\text{inj}}$  by less than  $1\sigma$ , and (ii)  $k_{\text{est}}$  differs from 0 by at least  $3\sigma$ . Using this criterion, about 45 of planets with RV semi-amplitudes lower than  $1.0 \text{ m s}^{-1}$  are detected from the activity-filtered RVs. This detection rate slowly decreases until  $K_{\text{inj}} \approx 0.2 \text{ m s}^{-1}$ , where 29 per cent of the injected planets are still recovered, and drops near zero for lower values of  $K_{\text{inj}}$ . The planet detection rate decreases roughly linearly with the planet orbital period, and is about 30 per cent and (resp. 20 per cent) for planet with orbital periods greater than 100 d (resp. 300 d). The fraction of detected planets from  $v_{\text{obs}}$  and  $v_{\perp}$  is shown in the  $(K_{\text{inj}}, P_{\text{orb}})$  space in Fig. B1. We find that, on average, our sensitivity to planet signatures increases by about 20 per cent from  $v_{\text{obs}}$  to  $v_{\perp}$ , and that this increase is the most spectacular for Earth-mass planets with orbital periods larger than 100 d, where the detection rate skyrockets by  $\sim 50$  per cent. However, Earth-mass planets with orbital periods larger than 300 d remain mostly undetected in both time-series, which is in line with the predictions of Meunier et al. (2023).<sup>4</sup>

In order to assess how the sampling strategy affects our activity-filtering framework, we perform the planet injection-recovery for different numbers of randomly selected data points, between 250 and 1750. For each sampling, we repeat the injection-recovery of the same 1000 planet signals (using the same sampling for all planets). Fig. 9 shows the mean error  $K_{\text{err}}$  on the RV semi-amplitude recovered with and without activity filtering. On average,  $K_{\text{err}}$  is reduced by a factor  $\sim 2$  between  $v_{\text{obs}}$  and  $v_{\perp}$ . In both cases, the evolution of the detection rate with the number of points is well described by a square-root law (with reduced  $\chi^2$  of 1.0 and 1.1 for  $v_{\text{obs}}$  and  $v_{\perp}$ ; see Fig. 9). This suggests that the precision on  $k_{\text{est}}$  is mostly driven by the number of data points used in the fit, despite the presence of supergranulation, activity residuals and instrumental systematics.

<sup>4</sup>Note also that, the evolution of the planet detection rate from  $v_{\perp}$  with the planet mass is found roughly similar to that shown, for example, in the fig. 8 of Meunier et al. (2023)

## 5 GAUSSIAN PROCESS REGRESSION

Our activity-filtering framework provides a new time series of RVs,  $v_{\perp}$ , less dispersed than the input RVs,  $v_{\text{obs}}$ . In the last section, we demonstrated that planet signatures in  $v_{\text{obs}}$  would be largely preserved in  $v_{\perp}$ . However, while the activity-filtering algorithm turns out to be efficient at filtering long-term activity effects, it leaves a significant rotationally modulated component in  $v_{\perp}$ , which also needs to be filtered out in order to improve the detectability of small planet signatures. Gaussian Process regression has become widely used to model quasi-periodic signals (Haywood et al. 2014; Rajpaul et al. 2015; Aigrain & Foreman-Mackey 2022; Nicholson & Aigrain 2022; Barragán et al. 2022b). In this section, we investigate how GPs can be used to model the stellar activity contributions in  $v_{\text{obs}}$  and  $v_{\perp}$ . In Section 5.1, we first investigate how the statistical properties of  $v_{\text{obs}}$  and usual stellar activity indicators compare to each other and evolve over time. In Section 5.2, we assess the ability of GPs to model the quasi-periodic activity signals in  $v_{\perp}$ , and in  $U_1$  to  $U_4$ . In Section 5.3, we finally investigate how a joint multidimensional GP-based activity filtering and planet search algorithm will perform for different indicators and input RV signals. All GP modellings in this section were performed using the open-source software PYANETI (Barragán, Gandolfi & Antoniciello 2019; Barragán et al. 2022a) with non-informative priors for all parameters.

### 5.1 Modelling of usual activity proxies

We start by independently modelling the HARPS-N RV time-series,  $v_{\text{obs}}$ , as well as the time series of usual activity proxies (FWHM,  $V_s$ ,  $S_{\text{HK}}$ ), using a one-dimensional (1D) GP with quasi-periodic covariance kernel  $k$  between each pair of epochs  $t_i$  and  $t_j$  given by

$$k(t_i, t_j) = A^2 \exp \left[ -\frac{\sin^2(\pi(t_j - t_i)/P_{\text{GP}})}{2\lambda_p} - \frac{(t_j - t_i)^2}{2\lambda_e} \right], \quad (6)$$

where  $P_{\text{GP}}$  is the GP period, which corresponds to the stellar rotation period (Angus et al. 2018; Nicholson & Aigrain 2022),  $\lambda_p$  is the inverse harmonic complexity and  $\lambda_e$  is the GP evolution time-scale (already defined in equation 4). Each time-series is modelled by six parameters: three hyper-parameters in the covariance matrix (equation 6), the GP amplitude  $A$ , a constant offset, and one additional uncorrelated jitter term  $\sigma_j$ , added in quadrature to the diagonal of the covariance matrix, to absorb any variations not accounted for by the GP. The parameter space is explored using the Bayesian Markov chain Monte Carlo (MCMC) sampler defined in Barragán et al. (2019), with uninformative priors for the parameters. To ensure that the MCMC process has converged, PYANETI iteratively runs 250 independent chains of 5000 steps until the chains converge, which is assessed using the statistics of Gelman & Rubin (1992). As detailed in the section 2.5 of Barragán et al. (2019), the code compares the variances between chains and within chains and uses the scaling factor  $\hat{R}$  introduced in Gelman et al. (2004) to assess convergence (typically with  $\hat{R} < 1.02$ ). We then run a last MCMC process with 5000 steps of 250 independent chains, corresponding to 125 000 independent samples for each parameter, from which we estimate the best-fitting value with  $1\sigma$  uncertainties.

Since the HARPS-N solar data set covers more than half of the Sun's activity cycle, it is likely that the statistical properties (i.e. evolution time scale, harmonic complexity, period) of the activity RV signal vary over the course of the observations. We therefore independently analyse each of the three seasons defined in Table 2 (i.e. 2015–2018, 2018–2021, 2021–2023). The best-fitting GP hyper-

**Table 4.** Best-fitting evolution time scale ( $\lambda_e$ ), inverse harmonic complexity ( $\lambda_p$ ) and period ( $P_{GP}$ ) of the one-dimension GP fit to the different time series analysed in this study.

	End Cycle 24: 2015–2018			Solar minimum: 2018–2022			Start Cycle 25: 2022–2024		
	$\lambda_e$ [d]	$\lambda_p$	$P_{GP}$ [d]	$\lambda_e$ [d]	$\lambda_p$	$P_{GP}$ [d]	$\lambda_e$ [d]	$\lambda_p$	$P_{GP}$ [d]
$v_{obs}$	$18.0 \pm 1.6$	$0.22 \pm 0.02$	$26.37^{+0.45}_{-0.44}$	$21.8 \pm 1.5$	$0.32 \pm 0.03$	$27.25 \pm 0.35$	$21.4^{+1.3}_{-1.4}$	$0.32 \pm 0.02$	$27.39 \pm 0.36$
FWHM	$22.2 \pm 1.1$	$0.51 \pm 0.03$	$27.70^{+0.43}_{-0.42}$	$27.7^{+1.8}_{-1.7}$	$0.54 \pm 0.04$	$27.58^{+0.37}_{-0.35}$	$22.9^{+1.2}_{-1.3}$	$0.56 \pm 0.04$	$27.59^{+0.43}_{-0.44}$
$V_s$	$19.8 \pm 1.1$	$0.28 \pm 0.02$	$26.43^{+0.37}_{-0.33}$	$22.7^{+1.6}_{-1.5}$	$0.33 \pm 0.02$	$27.16 \pm 0.35$	$23.0^{+1.2}_{-1.3}$	$0.32 \pm 0.02$	$27.45^{+0.28}_{-0.27}$
$S_{HK}$	$21.6 \pm 1.2$	$0.59 \pm 0.04$	$28.01^{+0.49}_{-0.51}$	$23.9^{+1.3}_{-1.4}$	$0.55 \pm 0.04$	$27.81^{+0.44}_{-0.45}$	$22.7^{+1.3}_{-1.4}$	$0.57^{+0.04}_{-0.03}$	$28.75^{+0.46}_{-0.45}$
$v_{  }$	$22.1 \pm 1.5$	$0.52 \pm 0.04$	$28.75 \pm 0.62$	$25.5^{+2.5}_{-2.3}$	$0.70^{+0.10}_{-0.08}$	$25.67^{+0.66}_{-0.64}$	$22.3^{+1.3}_{-1.4}$	$0.60^{+0.05}_{-0.04}$	$27.63^{+0.50}_{-0.48}$
$U_1$	$22.2 \pm 1.4$	$0.55 \pm 0.04$	$28.65^{+0.52}_{-0.53}$	$26.8^{+2.4}_{-2.2}$	$0.67^{+0.07}_{-0.06}$	$26.23^{+0.56}_{-0.57}$	$22.2 \pm 1.5$	$0.59 \pm 0.04$	$28.47^{+0.51}_{-0.59}$
$U_3$	$22.7^{+4.1}_{-2.9}$	$0.21 \pm 0.04$	$26.11^{+0.50}_{-0.40}$	$20.7 \pm 3.8$	$0.49^{+0.28}_{-0.13}$	$26.58^{+1.88}_{-1.21}$	$18.2 \pm 2.3$	$0.21^{+0.03}_{-0.02}$	$27.55^{+0.51}_{-0.49}$
$U_4$	$23.0^{+2.8}_{-2.4}$	$0.20 \pm 0.02$	$26.75^{+0.31}_{-0.29}$	$23.4^{+5.6}_{-4.3}$	$0.32^{+0.09}_{-0.06}$	$26.93^{+1.15}_{-0.83}$	$24.1^{+2.6}_{-2.2}$	$0.21 \pm 0.02$	$26.99^{+0.24}_{-0.23}$
$v_{\perp}$	$18.8^{+3.7}_{-2.4}$	$0.18^{+0.03}_{-0.02}$	$26.01^{+0.58}_{-0.54}$	$20.7^{+1.9}_{-1.8}$	$0.35^{+0.05}_{-0.04}$	$27.32^{+0.62}_{-0.59}$	$21.9^{+2.2}_{-2.3}$	$0.23 \pm 0.02$	$27.36^{+0.38}_{-0.43}$

parameters for each season are given in Table 4, and their associated 1D posterior distributions are shown in Fig. 10.

Significant variations are observed from one season to the next. The RV GP period increases from  $\sim 26.4$  d, in the 2015–2018 season, to  $\sim 27.4$  d after 2022 (i.e.  $\sim 2\sigma$  increase). This is in line with expectations as active regions are found at higher latitudes in the start of the activity cycle and migrate near the equator at the end of the cycle (see Hathaway 2010, for a review of solar activity cycle). The difference in rotation periods then simply reflects the latitudinal differential rotation of the Sun (e.g. Howard & Harvey 1970; Snodgrass & Ulrich 1990; Beck 2000). For example, the average latitudes of sunspots computed from the Solar Optical Observing Network of the US Air Force (USAF), with the help of the US National Oceanic and Atmospheric Administration (NOAA), is about  $12^\circ$  in 2015–2018 and  $20^\circ$  in 2022–2024<sup>5</sup>, corresponding to a difference of about 1d in rotation period (using the latitudinal differential rotation profile of Snodgrass & Ulrich 1990), consistent with our estimates in Table 4.

We also note that the RV GP evolution time scales ( $\lambda_e$ ) are almost  $2\sigma$  larger (i.e.  $\sim 3$  d) past 2018. Disc-resolved solar observations found larger active regions in the 2022–2024 period than in 2015–2018<sup>5</sup>, which is consistent with our observations as the active region evolution time-scale roughly scales with the surface area of the features (e.g. Leighton 1964; Wang, Nash & Sheeley 1989; Foukal 1998). The higher harmonic complexity (i.e. lower  $\lambda_p$ ) in the 2015–2018 season can be explained by the fact that equatorial active regions perturbs the wings of the profile more than higher-latitude regions and thus have an higher impact on the flux derivative which induce a sharper RV signature. As expected,  $\lambda_p$  is about twice as large for indicators of the photospheric flux (FWHM,  $S_{HK}$ ) than for indicators of the first temporal derivative of the flux, like RV  $V_s$  (e.g. Aigrain, Pont & Zucker 2012; Klein & Donati 2019; Nicholson & Aigrain 2022). The evolution of the GP covariance kernel over time, and its effect on the modelling of RV time-series, is discussed in Section 6.1.

## 5.2 Modelling activity-filtered RVs with Gaussian Processes

Using the 1D GP framework described in Section 5.1, we model the time-series of  $v_{\perp}$ ,  $v_{||}$ , as well as  $U_1$  to  $U_4$ . Except  $U_2$ , for which the MCMC process does not converge, all time-series exhibit a quasi-periodic modulation well captured by the GP. For these time-series,

the best-fitting GP hyper-parameters are given in Table 4 and the posterior distributions of the fit parameters are shown for  $v_{||}$  and  $v_{\perp}$  in Fig. 11.

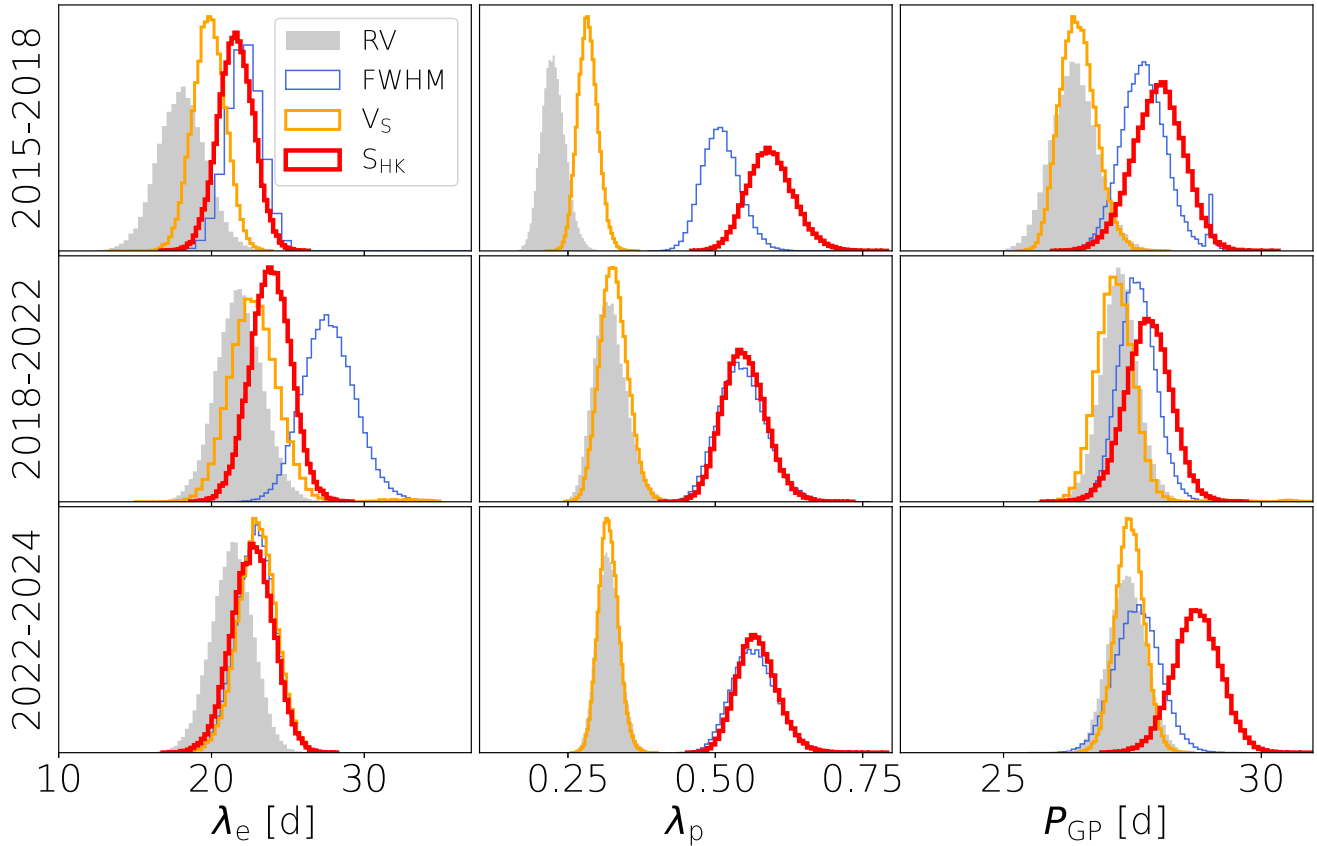
As already intuited from Fig. 6, we find that the shape-driven RVs,  $v_{||}$ , are systematically smoother (i.e. with larger values of  $\lambda_p$ ) than  $v_{obs}$ , which suggests that the activity signal within  $v_{||}$  is mostly driven by variations of the photospheric flux (typically induced by faculae and plages). This is further evidenced by the fact that  $v_{||}$  tends, in general, to evolve on longer time-scales than  $v_{obs}$ . On the other hand, the shift-driven RVs  $v_{\perp}$ , exhibit a higher harmonic complexity and evolve on time-scales similar to  $v_{obs}$ , which suggests that the stellar activity signal in  $v_{\perp}$  is sensitive to the variations of the first derivative of the photospheric flux (e.g. induced by spots). As expected, the best-fitting GP amplitudes are consistent with the RMS of the different time-series given in Table 2.

In all three seasons, a value of  $0.25$ – $0.3$   $m s^{-1}$  is obtained for jitter term in  $v_{||}$ , much lower than its counterpart in  $v_{obs}$ . This is expected as, by definition,  $v_{||}$  contains only a small fraction of the white noise of HARPS RV and therefore the estimate of  $\sigma_j$  reflects more what the GP cannot model than the real dispersion of white noise in  $v_{||}$ . On the other hand, jitter terms of  $0.5$ – $0.6$   $m s^{-1}$  are found for  $v_{obs}$  and  $v_{\perp}$ <sup>6</sup>. In each case, the jitter term is considerably larger than the photon noise of  $\sim 0.1$   $m s^{-1}$ , on the daily binned RVs, which suggests that a significant fraction of the RV variation can be explained neither by the GP nor by formal uncertainties. In our daily binned data set, both oscillation- and granulation-driven RV variations are expected to be averaged out due to their short evolution time scales (e.g. Dumusque et al. 2011; Chaplin et al. 2019). In contrast, supergranulation, which evolves on time-scale of days, will be only partially reduced by our binning process. Recently, Al Moulla et al. (2023) and Lakeland et al. (2024) measured RV RMS contributions of  $\sim 0.7$   $m s^{-1}$  and  $\sim 0.9$   $m s^{-1}$  for supergranulation in the HARPS-N solar data, consistent with the simulations of Meunier et al. (2015). Therefore, the dispersion budget in the RV residuals is likely a mix of supergranulation and long-term stability ( $\sim 0.4$   $m s^{-1}$  for HARPS-N).

The posterior densities of the hyper-parameters of the GP fit to  $U_1$ ,  $U_3$ , and  $U_4$  are shown in Fig. C4, and their best-fitting values are given in Table 4. As  $U_1$  is by far the dominant component of the SVD to the shape-driven CCFs, it follows a very similar behaviour to  $v_{||}$  and to  $S_{HK}$  (as expected from Fig. 5). On the other hand,  $U_3$  and  $U_4$  exhibit similar statistical properties as  $v_{\perp}$ , which suggests that they

<sup>5</sup><http://solarcyclescience.com/index.html>

<sup>6</sup>The typical error bar on the jitter term is about  $0.04$   $m s^{-1}$ .



**Figure 10.** One-dimensional posterior densities of the hyper-parameters of the GP fit to the HARPS-N solar RVs (filled grey histograms), and to the time series of FWHM (thin blue lines), bisector velocity span (gold lines), and  $S_{\text{HK}}$  (thick red lines), for the three seasons listed in Table 2. Columns 1, 2 and 3 show the posterior densities of the GP evolution time scale ( $\lambda_e$ ), inverse harmonic complexity ( $\lambda_p$ ) and period ( $P_{\text{GP}}$ ).

could be good proxies of the stellar activity signal in the latter time series. This is investigated in the next section.

### 5.3 Multidimensional GP analysis

In the practical case of the search for planet signatures in observed RVs, the shape-driven RVs,  $v_{\parallel}$  could potentially be used as planet-independent proxies to mitigate stellar activity signals. The multidimensional GP framework of Rajpaul et al. (2015) is one of the most robust ways to simultaneously model RVs and activity proxies, thereby boosting the number of constraints on the stellar activity parameters. On the other hand, the shift-driven RVs,  $v_{\perp}$ , are less dispersed than  $v_{\text{obs}}$ , which could boost the sensitivity to low-amplitudes planet signatures. Furthermore, we have shown in Section 5.2 that the residual quasi-periodic activity signals in  $v_{\perp}$  would be well modelled by a GP with quasi-periodic kernel. In this section, we explore how  $v_{\text{obs}}$  and  $v_{\perp}$  perform in the search for long-period low-amplitude planet signatures.

#### 5.3.1 Method

In our model, we assume that the activity signal in the RVs and indicators follow a FF'-like relation (Aigrain et al. 2012; Rajpaul et al. 2015). Each time-series  $\alpha$  is expressed as a linear combination of a latent variable  $G$  (typically the square of the photospheric flux)

and its first temporal derivative  $\dot{G}$ , such that, at time  $t$ ,

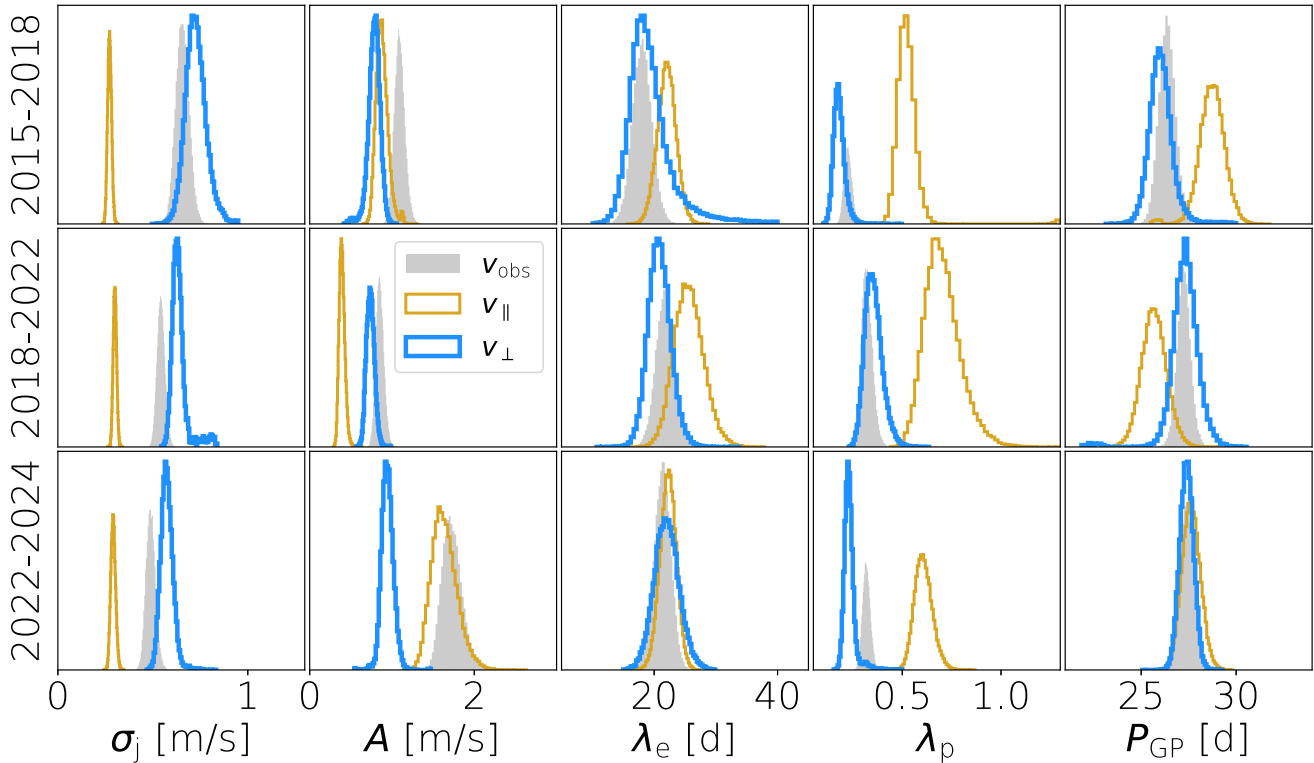
$$\alpha(t) = A_{\alpha}G(t) + B_{\alpha}\dot{G}(t) + C_{\alpha}, \quad (7)$$

where the amplitudes  $A_{\alpha}$ ,  $B_{\alpha}$ , and  $C_{\alpha}$  are free parameters of the model. The latent variable  $G$  is modelled as a GP with the quasi-periodic covariance kernel defined in equation (6).

We inject the RV signature of a single planet on a circular orbit to the HARPS-N solar CCFs. We assume an orbital period of 100 d and a RV semi-amplitude of  $0.4 \text{ m s}^{-1}$ , which corresponds to a planet mass of  $2.9 M_{\oplus}$ . In order to limit bias, three different orbital phases ( $\phi_p$  in equation 3) of 0.0, 0.3, and 0.7 are considered. We apply the framework of Section 2.1 to generate time series of shift- and shape-driven RVs, from the input RVs. We then use the multidimensional GP framework to model the activity signal within (1)  $v_{\text{obs}}$  and FWHM, (2)  $v_{\text{obs}}$  and  $V_s$ , (3)  $v_{\text{obs}}$  and  $S_{\text{HK}}$ , (4)  $v_{\text{obs}}$  and  $v_{\parallel}$ . These indicators exhibit relatively similar evolution time-scales and period to  $v_{\text{obs}}$ , and are therefore promising candidates for multi-dimensional GP modelling. Differences in  $\lambda_p$  are naturally accounted for in the framework of Barragán et al. (2022a).

Since we have demonstrated that planet signatures are mostly preserved in  $v_{\perp}$ , the planet search can as well be performed directly from this time series.<sup>7</sup> We therefore consider four additional cases, namely (5)  $v_{\perp}$  and FWHM, (6)  $v_{\perp}$  and  $V_s$ , (7)  $v_{\perp}$  and  $S_{\text{HK}}$ , and (8)  $v_{\perp}$  and  $v_{\parallel}$ . As a reference, we also model the stellar activity RV signals in  $v_{\text{obs}}$  (case  $R_0$ ) and in  $v_{\perp}$  (case  $R_{\perp}$ ) using uni-dimensional GPs with QP and SE kernels, respectively. In all cases, the GP is





**Figure 11.** Uni-dimensional posterior densities of the jitter term, GP amplitude and GP hyper-parameters of the fit to the HARPS-N solar RVs (filled grey distributions), the shape-driven RVs  $v_{\parallel}$  (thin gold lines), and the shift-driven RVs  $v_{\perp}$  (thick blue lines).

jointly modelled with a planet RV signature of fixed orbital period and phase, to simulate the RV monitoring of a transiting planet (e.g. unveiled by the PLATO space mission). Note also that, in all cases, the same planetary signal has been injected to the data, but the GP model differs from one case to the next. As shown in Section 5.1, the statistical properties of the time series vary significantly from one season to the next. We therefore choose to perform the planet injection-recovery tests independently on each of the three seasons defined in Table 2. For each season, our input data sets contain the same amount of data points (460), spread evenly over a  $\sim 2$ -yr period. The parameter space is sampled using the procedure described in Section 5.1.

### 5.3.2 Results

The best estimates of the planet RV semi-amplitude and of the jitter term  $\sigma_j$  are given in Table 5 and shown in Fig. 12. The recovered values of  $k_{\text{est}}$  are consistent with the injected one in all cases (note that, in all cases, the same planetary signal has been injected to the data), but the precision of the estimate varies significantly from one case to the next. Unsurprisingly, the 1D GP modelling of the HARPS-N RVs performs the least well, especially when the Sun is more active, in which case only a marginal  $\sim 2\sigma$  planet detection can be claimed. This is due to the fact that the long-term activity evolution and the planet signature cannot be easily separated with a simple 1D GP, which leads to larger error bars on  $k_{\text{est}}$ .

We find that multidimensional GPs give, in most cases, much smaller error bars on  $k_{\text{est}}$ . In particular, cases (1) and (3), where the FWHM of the CCF and the S index are used as activity proxies of  $v_{\text{obs}}$ ,

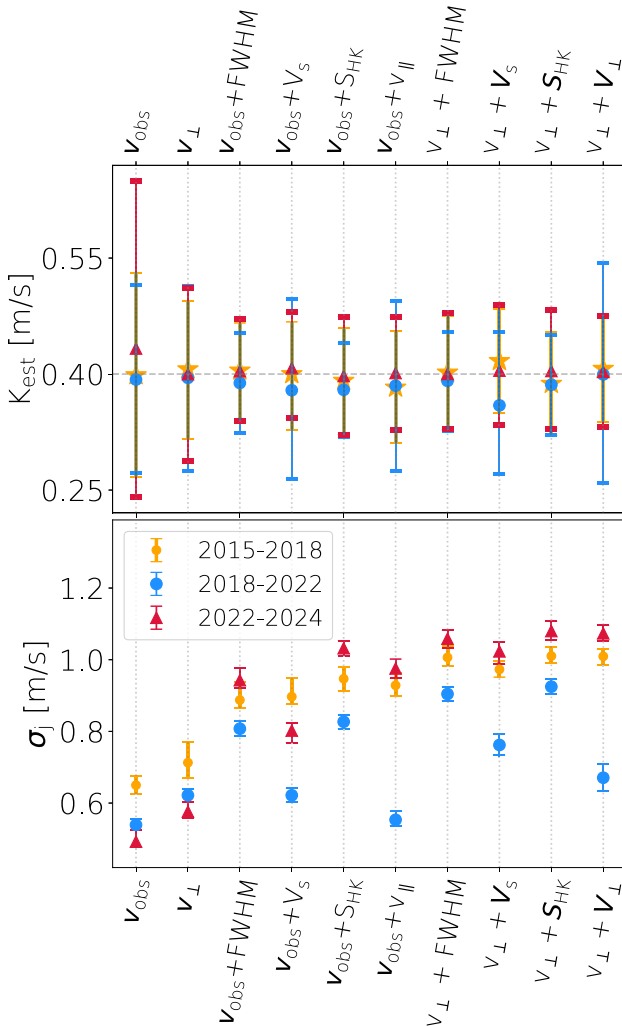
perform best in all three seasons. These two indicators capture well the activity signatures sensitive to the photospheric flux, and thereby complement well the RV signals, sensitive to variations of the flux and its first derivative. In these two cases, the multidimensional GP leverages all the constraints available, which results in more precise estimates of  $k_{\text{est}}$ . This also leads to higher values of  $\sigma_j$ , since the GP is now less flexible (due to the increased number of constraints), and less likely to absorb non-activity-induced variations.

We also note that cases (2) and (4), where  $v_{\text{obs}}$  is combined with  $V_s$  and  $v_{\parallel}$ , perform well when the Sun is more active but yield significantly larger error bars on  $k_{\text{est}}$  during solar minimum. During this period, there are fewer and smaller active regions than during more active epochs. It is therefore more likely that the statistical properties of the stellar activity RV signal vary significantly from one solar rotation to the next, and are thereby more difficult to model with a quasi-periodic GP (see Section 6.1). During solar minimum, neither  $V_s$  nor  $v_{\parallel}$  bring any additional constraint on the quasi-periodic activity RV signal and therefore the multidimensional GP modelling leads to results similar as the 1D GP modelling of  $v_{\text{obs}}$ .

More precise values of  $k_{\text{est}}$  are obtained in case  $R_{\perp}$  than in case  $R_0$ , which just reflects the fact that activity has been partly filtered from  $v_{\perp}$ . However, these values are significantly less precise than those obtained in the multidimensional GP framework, even when the computation of  $v_{\perp}$  is bypassed (i.e. cases 1 to 4). Cases where multidimensional GPs are used with  $v_{\perp}$  (i.e. cases 5 to 8) tend to perform similar as their  $v_{\text{obs}}$  counterpart (cases 1 to 4). In addition, we note that none of the SVD components,  $U_1$ ,  $U_3$ , and  $U_4$ , outperforms  $v_{\parallel}$  in the multidimensional GP framework. In particular, despite their periodic modulation,  $U_3$  and  $U_4$  do not improve the RV modelling compared to case  $R_0$ .

**Table 5.** Estimates of the planet RV semi-amplitude ( $k_{\text{est}}$ ) and RV jitter ( $\sigma_j$ ) for the different cases considered in Section 5.3. Note that, the semi-amplitude of the planet RV signature injected in the data is  $K_{\text{inj}} = 0.4 \text{ m s}^{-1}$ .

Case	Time series	2015–2018		2018–2022		2022–2024	
		$k_{\text{est}}[\text{m s}^{-1}]$	$\sigma_j [\text{m s}^{-1}]$	$k_{\text{est}}[\text{m s}^{-1}]$	$\sigma_j [\text{m s}^{-1}]$	$k_{\text{est}}[\text{m s}^{-1}]$	$\sigma_j [\text{m s}^{-1}]$
(R <sub>0</sub> )	$\mathbf{v}_{\text{obs}}$	$0.40 \pm 0.13$	$0.65^{+0.03}_{-0.02}$	$0.39 \pm 0.12$	$0.54 \pm 0.02$	$0.43^{+0.22}_{-0.19}$	$0.49^{+0.03}_{-0.02}$
(R <sub>⊥</sub> )	$\mathbf{v}_{\perp}$	$0.41 \pm 0.09$	$0.71^{+0.06}_{-0.04}$	$0.40 \pm 0.12$	$0.62 \pm 0.02$	$0.40 \pm 0.11$	$0.58^{+0.03}_{-0.02}$
(1)	$\mathbf{v}_{\text{obs}} + \text{FWHM}$	$0.40 \pm 0.06$	$0.89^{+0.05}_{-0.02}$	$0.39^{+0.07}_{-0.06}$	$0.81 \pm 0.02$	$0.40^{+0.07}_{-0.06}$	$0.94^{+0.03}_{-0.02}$
(2)	$\mathbf{v}_{\text{obs}} + \mathbf{V}_s$	$0.40 \pm 0.07$	$0.90^{+0.05}_{-0.02}$	$0.38^{+0.12}_{-0.11}$	$0.62 \pm 0.02$	$0.41 \pm 0.07$	$0.80^{+0.02}_{-0.03}$
(3)	$\mathbf{v}_{\text{obs}} + \mathbf{S}_{\text{HK}}$	$0.39 \pm 0.07$	$0.95 \pm 0.03$	$0.38 \pm 0.06$	$0.83 \pm 0.02$	$0.40 \pm 0.08$	$1.03 \pm 0.02$
(4)	$\mathbf{v}_{\text{obs}} + \mathbf{v}_{\perp}$	$0.38 \pm 0.07$	$0.93 \pm 0.03$	$0.38 \pm 0.11$	$0.55 \pm 0.02$	$0.40 \pm 0.07$	$0.98 \pm 0.03$
(5)	$\mathbf{v}_{\perp} + \text{FWHM}$	$0.40 \pm 0.07$	$1.01^{+0.03}_{-0.02}$	$0.39 \pm 0.06$	$0.90 \pm 0.02$	$0.40^{+0.08}_{-0.07}$	$1.06^{+0.02}_{-0.03}$
(6)	$\mathbf{v}_{\perp} + \mathbf{V}_s$	$0.42 \pm 0.07$	$0.97 \pm 0.02$	$0.36 \pm 0.09$	$0.76 \pm 0.03$	$0.40^{+0.09}_{-0.07}$	$1.02^{+0.03}_{-0.04}$
(7)	$\mathbf{v}_{\perp} + \mathbf{S}_{\text{HK}}$	$0.39 \pm 0.07$	$1.01^{+0.03}_{-0.02}$	$0.39^{+0.06}_{-0.07}$	$0.92 \pm 0.02$	$0.40^{+0.08}_{-0.07}$	$1.08^{+0.03}_{-0.02}$
(8)	$\mathbf{v}_{\perp} + \mathbf{v}_{\parallel}$	$0.41 \pm 0.07$	$1.01 \pm 0.02$	$0.40 \pm 0.14$	$0.67 \pm 0.04$	$0.40 \pm 0.07$	$1.07 \pm 0.02$

**Figure 12.** Best estimates of the planet RV semi-amplitude ( $k_{\text{est}}$ , top panel) and jitter term ( $\sigma_j$ , bottom panel) for the different cases listed in Table 5. The orange dots, light blue filled circles, and red triangles indicate the results obtained for the 2015–2018, 2018–2022, and 2022–2024 season, respectively. The dashed horizontal line on the top panel indicates the semi-amplitude of the planet RV signature injected in the data.

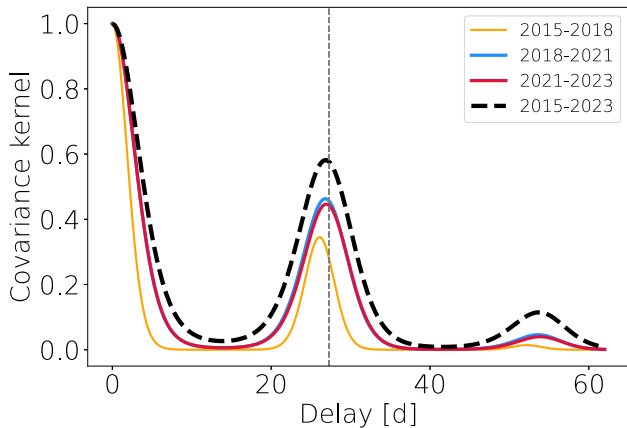
## 6 DISCUSSION AND CONCLUSIONS

In this paper, we analysed the activity-induced distortions in the absorption lines of the Solar spectrum, intensively monitored with HARPS-N over the last 8 yr. From the DRS-processed systematic-corrected CCFs, we constructed a dimensionally reduced basis, which allowed us to separate the observed RVs ( $\mathbf{v}_{\text{obs}}$ ) into complementary time-series, namely  $\mathbf{v}_{\perp}$  (i.e. shift-driven RVs) and  $\mathbf{v}_{\parallel}$  (shape-driven RVs). In the first one, the variations are largely dominated by Doppler shifts of the entire CCF, induced, for example, by planets or granulation at the solar surface. In the second one, the variations are only driven by CCF distortions and, thereby, probe the effects of active regions whilst being independent of planet signatures. When we apply this framework to the HARPS-N solar spectra, the RMS of the observed RVs goes from 2.05 to 1.06  $\text{m s}^{-1}$ , hence a decrease of about 50 percent. We find that planet signatures are mostly preserved in  $\mathbf{v}_{\perp}$ , and that the efficiency of the activity filtering is not affected by the temporal sampling. However, as shown in Fig. 7,  $\mathbf{v}_{\perp}$  exhibits significant rotation-induced variations, suggesting that the shift-driven RVs are still affected by stellar activity.

### 6.1 Evolution of the GP covariance structure

In order to better understand the nature of the rotational modulation within  $\mathbf{v}_{\text{obs}}$ ,  $\mathbf{v}_{\perp}$ , and  $\mathbf{v}_{\parallel}$ , all time-series were modelled using a GP with quasi-periodic covariance kernel. As shown in Fig. 13, the structure of the covariance matrix of the time series evolves significantly over time. The RV variations are smoother and more slowly evolving at the start of Cycle 25 than at the end of Cycle 24, which reflects the fact that the Sun’s active regions are larger and at higher-latitude in the start of the magnetic cycle. We found that the quasi-periodic RV activity signals driven by flux variations as well as the long term cycle evolution are well captured by  $\mathbf{v}_{\parallel}$ . On the other hand,  $\mathbf{v}_{\perp}$  is still plagued by high-complexity quasi-periodic activity signals, likely reflecting variations in the first temporal derivative of the photospheric flux.

To visualize how the temporal evolution of the GP covariance kernel affects our GP fit, we model the full HARPS-N RV time-series with a single 1D GP, using the quasi-periodic covariance kernel of equation (6). The best-fitting covariance kernel, shown in dashed black line in Fig. 13, differs significantly from the kernels obtained individually for the different seasons. The signal is now found smoother ( $\lambda_p \approx 0.4$ ) and more slowly evolving ( $\lambda_e \approx 26$  d). On the other hand, the uncorrelated jitter term  $\sigma_j$  is now  $\sim 1.8 \text{ m s}^{-1}$ ,

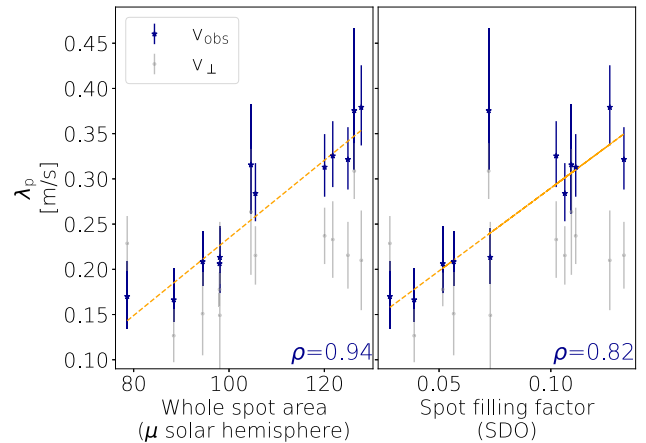


**Figure 13.** Evolution of the covariance kernel of the stellar activity RV signal, computed using equation (6) with the best-fitting hyperparameters of the GP modelling. The thick dashed line is obtained by modelling all RVs together and the vertical dotted line indicate the average synodic rotation period of the Sun (27.2753 d).

almost three times greater than on individual seasons. As explained in Section 5.3, we expect  $\sigma_j$  to increase, since we have tripled the number of constraints in the fit. On the other hand, the values of  $\sigma_j$  obtained with the multidimensional GP in Table 5 are about  $0.8\text{--}1.0\text{ m s}^{-1}$ , significantly smaller than  $1.8\text{ m s}^{-1}$ . Moreover, such a high value can be explained neither by supergranulation or long-term instrument stability. We therefore conclude that, as things stand, quasi-periodic stellar activity RV signals are better modelled by GPs on chunks during which cycle-driven variations remain small.

If we now model the full time series of shift-driven RVs,  $v_{\perp}$ , with a single 1D GP, the best-fitting hyper-parameters are fully consistent with those obtained on individual seasons, despite the variations of  $P_{\text{GP}}$  between 2015 and 2024. The best-fitting evolution time-scale and GP period lie in between the estimates listed in Table 4, and a model with high harmonic complexity ( $\lambda_p = 0.23 \pm 0.01$ ) is preferred. The value of  $\sigma_j$  ( $0.62 \pm 0.02\text{ m s}^{-1}$ ) lies in between the values obtained on individual seasons (between  $0.57$  and  $0.72\text{ m s}^{-1}$ ), which indicates that our fit is as good on the whole data set as it is on individual seasons. This can be explained by the fact that most of the long-term activity variations (e.g. induced by the magnetic cycle), are well captured by  $v_{\parallel}$  and, thus, well filtered out from  $v_{\perp}$ . Therefore, we expect the GP covariance matrix to exhibit fewer variations in the case of  $v_{\perp}$  than in the case of  $v_{\text{obs}}$ .

To investigate what drives the evolution of the quasi-periodic covariance kernel in  $v_{\text{obs}}$ , we divide the HARPS-N RVs into 270-d chunks (i.e. 10 rotation periods), using a moving window with a step of 100 d. After discarding chunks containing large gaps in the data, we model each subset of data using a GP with the quasi-periodic kernel given in equation (6). The best-fitting hyper-parameters are shown in Fig. C5. The variation of the GP amplitude, which is well described by a parabola, is unsurprisingly a good tracer of the solar magnetic cycle. We do not observe significant variations either in the time-scale of the evolution of the GP or in the period, due to the relatively large uncertainties over these two parameters (about 4 d and 1 d for  $\lambda_c$  and  $P_{\text{GP}}$ , respectively). On the other hand, inverse harmonic complexity,  $\lambda_p$ , varies significantly from one chunk to the next. When discarding the values obtained during solar minimum, often plagued with large uncertainties due to the weak activity signal, we find a positive correlation with the GP amplitude (Pearson correlation coefficient of 0.7).



**Figure 14.** GP inverse harmonic complexity extracted from the HARPS-N RVs ( $v_{\text{obs}}$ , dark blue stars) as a function of the whole sunspot area (in millionths of solar hemisphere; left-hand panel), and of the SDO spot filling factor (right-hand panel). In each panel, we give the Pearson correlation coefficient  $\rho$ , and show the best-fitting straight line (orange dashed lines). For comparison, we also show the inverse harmonic complexities extracted from the shift-driven RVs ( $v_{\perp}$ ) in grey dots, with Pearson correlation coefficients of 0.55 and 0.39 with the whole spot area and SDO spot filling factor, respectively.

In order to better understand the origin of this correlation, we compare in Fig. 14 the values of  $\lambda_p$  to the whole sunspot area observed by USAF/NOAA<sup>7</sup>, averaged over each chunk outside the solar minimum, and to the spot filling factor derived from SDO HMI observations (using the method of Milbourne et al. 2019; Ervin et al. 2022; Haywood et al. 2022; Lakeland et al. 2024). Both quantities appear to strongly correlate with  $\lambda_p$  (Pearson correlation coefficients of 0.94 and 0.82; see Fig. 14). When the spot filling factor is large, sunspots are likely to be more evenly distributed in longitude at the stellar surface, which results in a smoother RV signal (i.e. greater values of  $\lambda_p$ ). Conversely, for small filling factors, sunspots are more likely to be longitudinally isolated and create sharper RV variations. We also find that  $\lambda_p$  exhibits a weaker correlation with the filling factor of faculae and plages (Pearson correlation coefficient of  $\sim 0.6$ ). This suggests that long-term variations in  $\lambda_p$  are due more to sunspots than to the magnetic cycle, known to be best described by the filling factor of faculae/plages over the course of our observations (Cretignier, Pietrow & Aigrain 2024). The inhibition of convective blueshift in faculae induces an RV contribution that evolves, to the first order, as the square of the disc-integrated photospheric flux (Aigrain et al. 2012). Therefore, in absence of spots, the RV activity signals should not vary less smoothly than the FWHM or the S index. On the other hand, spot-induced activity signals, which vary as the derivative of the photospheric flux, are characterized by a significantly higher harmonic complexity. Therefore, despite the fact that faculae dominate the stellar activity RV budget, spots remain the main driver of the smoothness of the signal, probed by  $\lambda_p$ . Despite the fact that this result has been intuited in the literature (e.g. Nicholson & Aigrain 2022), it is the first time that it is observed on data.

<sup>7</sup>Note that, as discussed in Section 2, small-amplitude planet signatures are expected to be preserved in the shift-driven RVs,  $v_{\perp}$ . Coupling a multi-dimensional GP framework with the likelihood model of equation (5) will be implemented in a forthcoming work.

In contrast to  $v_{\text{obs}}$ , no coherent variations in the statistical properties of  $v_{\perp}$  are observed. The GP amplitude remains roughly constant over time, and the correlation between  $\lambda_p$  and the spot filling factor (or whole spot area) is now much weaker (Pearson correlation coefficients of 0.55 and 0.39; see Fig. 14) is observed. This confirms that, unlike  $v_{\text{obs}}$ , the quasi-periodic activity signal in  $v_{\perp}$  can be modelled by a single GP on time scales of several years (potentially over the full cycle).

## 6.2 Impact of systematics

The analysis presented in this study was conducted on CCFs computed from spectra processed by YV1, as described in Section 3.2. Most of the corrected systematics are expected to impact the shape of the CCFs and therefore the activity-filtering framework described in Section 2.1. To quantify this impact, we repeat the analysis presented in Section 4.1 with the non-YARARA-processed spectra. The spectra selected in Section 3.1 are normalized and aligned using the procedure described in Section 3.2, and CCFs are computed using the ‘Kit-Cat’ line list of Cretignier et al. (2022). Following the method described in Section 2.1, we extract the time-series of  $v_{\parallel}$  and  $v_{\perp}$ , using the first 4 components in the basis projection.

We find that, whereas the RMS of  $v_{\text{obs}}$  remains pretty similar in the post-processed counterpart, the shift-driven RVs  $v_{\perp}$  (resp. shape-driven RVs  $v_{\parallel}$ ) are now significantly more (resp. less) dispersed, with a dispersion of  $1.36 \text{ m s}^{-1}$  RMS (resp.  $1.52 \text{ m s}^{-1}$  RMS). When working independently on the three seasons defined in Table 2, we only see a noticeable decrease in RV RMS in Season 3. Component  $U_1$  still correlates well with the FWHM of the CCF, but both time-series are dominated by 0.5- and 1-yr periodic modulations, induced by the Earth orbital obliquity and eccentricity (Collier Cameron et al. 2019). In contrast, no more correlation between  $U_1$  and  $v_{\text{obs}}$ ,  $V_s$  and  $S_{\text{HK}}$  is observed. Moderate correlations (Pearson correlation coefficients up to  $\sim 0.6$ ) are observed between these quantities and higher-order components, which reflects the fact that information has been diffused among the SVD components, due to systematics contributing as much, if not more, to the CCF variations.

When the RV time-series are modelled with a quasi-periodic 1D GP, we find that  $v_{\parallel}$  is significantly smoother and slowly-evolving than its YV1-processed counterpart. The fit is now poorer, with a residual RMS twice as large. This is due to the fact that  $v_{\parallel}$  is plagued with systematics (e.g. due the 0.5-yr and 1-yr variations in FWHM Collier Cameron et al. 2019), which affects the GP modelling. On the other hand, whereas no significant changes in the statistical properties of  $v_{\perp}$  are observed, the values of  $\sigma_j$ , the uncorrelated jitter term, have increased to  $\sim 1 \text{ m s}^{-1}$ , which means that the fit is poorer. We conclude that our framework to correct for shape-driven RV variations works best when systematics have been corrected beforehand, through a dedicated post-processing technique. The latter helps to better isolate stellar activity signals in the spectra, which can therefore be more accurately modelled.

## 6.3 Activity filtering and planet recovery

By jointly modelling the stellar activity RV signal in  $v_{\text{obs}}$  with different activity proxies, either extracted from the input spectra (e.g. CCF FWHM,  $S_{\text{HK}}$ ) or from our framework (i.e.  $v_{\parallel}$ ), we demonstrated that long-period planets with small RV semi-amplitudes could be reliably detected with a multidimensional GP framework (see Table 5). This experiment also confirms that modelling the reduced RVs with a one-dimensional GP will likely yield imprecise parameters for long-period planets, whereas multidimensional GPs, by increasing

the number of constraints on the fitted parameters, will, in most cases, provide a more robust treatment of long-term variations.

Further investigations are also needed to leverage all the information available in the wavelength space (CCFs<sup>8</sup>, spectra) to extract planet signatures in an optimal way and increase the constraints on their parameters (following the works of e.g. Dumusque 2018; Rajpaul et al. 2020; Collier Cameron et al. 2021; Al Moulla et al. 2022; Cretignier et al. 2022; de Beurs et al. 2022). In particular, methods like Doppler Imaging (Kochukhov 2016; Luger et al. 2021; Asensio Ramos, Díaz Baso & Kochukhov 2022; Klein et al. 2022), which model the full absorption line profile (or spectrum), thereby bypassing the computation of RVs, might become robust complements to usual activity modelling techniques. Finally, long-term variations of the activity properties over the cycle (as evidenced in Fig. 13) are still hard to reproduce by current state-of-the-art modelling tools like GPs. Defining more physically driven GP kernels (e.g. Hara & Delisle 2023) or allowing some hyper-parameters to vary with time are potential avenues for solving this problem.

Yet, even if the best-case scenarios, the RV residuals of the GP fit exhibit RMSs greater than  $\sim 0.5 \text{ m s}^{-1}$ , significantly larger than the formal RV uncertainties of  $\sim 0.1 \text{ m s}^{-1}$  for the daily binned data. This high dispersion is most likely attributable to the long-term instrument stability and to the solar supergranulation. This phenomenon, whose origin remains unclear (see Rincon & Rieutord 2018, for a review), is expected to induce RV signals on the same order of magnitude as our residuals (Meunier et al. 2015; Al Moulla et al. 2023). Moreover, as we do not expect granulation signals to have obvious effects on the shape of the spectral lines, the activity-filtering method presented in Section 2.1 would not correct for them. Similarly, our GP model is not expected to affect supergranulation signals, as it is designed to account for rotationally modulated activity signals which evolve on significantly longer time-scales than supergranulation. Averaging out the RV effect of supergranulation using dedicated sampling strategies, as generally done for oscillations- and granulation-induced RV signals (Dumusque et al. 2011), is not a straightforward option here, as it will require the star to be intensively observed on time scales of days. As indicated in Meunier & Lagrange (2019), better understanding the origin of supergranulation, and developing physically or data-driven methods to model its RV contributions (O’Sullivan & Aigrain 2024, see), will be an important step to detect Earth-mass planets as part, for example, of PLATO ground-based monitoring or dedicated RV surveys like the Terra Hunting Experiment (Hall et al. 2018). As things currently stand, supergranulation signals have to be treated as white noise, which is emphasized by the fact that, in Section 4.2, the precision of the recovered planet semi-amplitudes improves as the squared root of the number of measurements. If this result is confirmed by dedicated studies, it would mean that RV monitoring missions should aim at maximizing the number of measurements on each target (ensuring these are spaced at least 1–2 d apart so that the super-granulation signal is uncorrelated), even at the cost of monitoring fewer stars overall.

## ACKNOWLEDGEMENTS

We warmly thank the reviewer, Dr Drake Deming, for valuable suggestions which helped make this analysis clearer and more robust.

<sup>8</sup>Note that, CCFs may not be the most accurate way of estimating line profiles (in comparison to e.g. profiles obtained via least-squares deconvolution; see Donati & Brown 1997; Lienhard et al. 2022).



We also warmly thank A. A. John for pointing out confusing notations in early versions of the manuscript. The HARPS-N project has been funded by the Prodex Program of the Swiss Space Office (SSO), the Harvard University Origins of Life Initiative (HUOLI), the Scottish Universities Physics Alliance (SUPA), the University of Geneva, the Smithsonian Astrophysical Observatory (SAO), and the Italian National Astrophysical Institute (INAF), the University of St Andrews, Queen's University Belfast, and the University of Edinburgh. We thank the HARPS-N solar team and TNG staff for processing the solar data and maintaining the solar telescope. BK, SA, OB, HY, and NKOS acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no 865624, GPRV). MC acknowledges the SNSF support under grant P500PT\_211024. FR is funded by the University of Exeter's College of Engineering, Maths and Physical Sciences, UK. ACC acknowledges support from STFC consolidated grant number ST/V000861/1, and EPSRC grant number EP/Z000181/1 towards the ERC Synergy project REVEAL.

## DATA AVAILABILITY

This work makes use of the HARPS-N solar RVs, which will be described and made available in Dumusque et al., submitted. The SDO/HMI images are publicly available at <https://sdo.gsfc.nasa.gov/data/> and the USAF/NOAA observations are available at <http://solarcyclescience.com/index.html>.

## REFERENCES

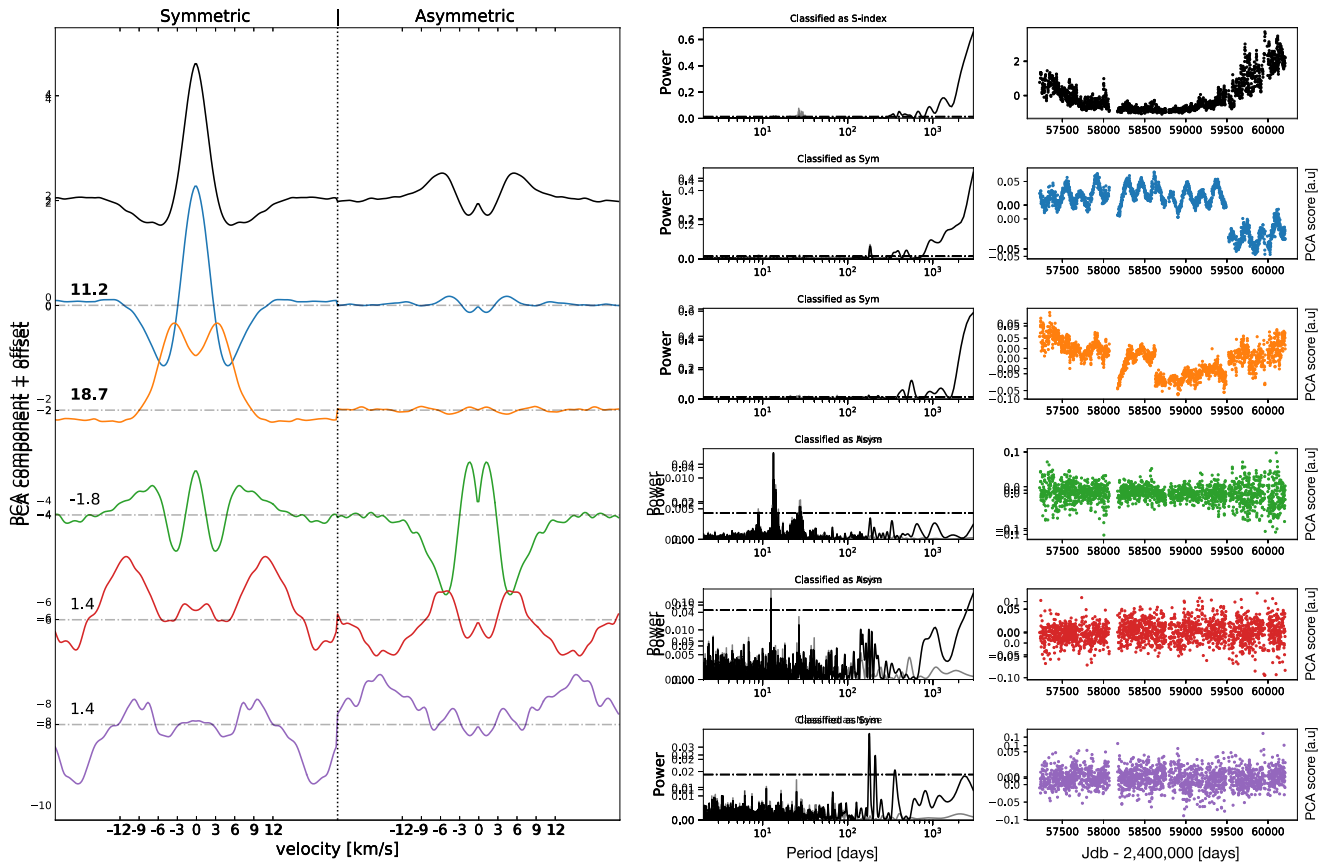
- Aigrain S., Foreman-Mackey D., 2022, *Annu. Rev. Astron. Astrophys.*, 61, 329
- Aigrain S., Pont F., Zucker S., 2012, *MNRAS*, 419, 3147
- Al Moulla K., Dumusque X., Cretignier M., Zhao Y., Valenti J. A., 2022, *A&A*, 664, A34
- Al Moulla K., Dumusque X., Figueira P., Lo Curto G., Santos N. C., Wildi F., 2023, *A&A*, 669, A39
- Ambikasaran S., Foreman-Mackey D., Greengard L., Hogg D. W., O'Neil M., 2015, *IEEE Trans. Pattern Anal. Mach. Intell.*, 38, 252
- Angus R., Morton T., Aigrain S., Foreman-Mackey D., Rajpaul V., 2018, *MNRAS*, 474, 2094
- Asensio Ramos A., Díaz Baso C. J., Kochukhov O., 2022, *A&A*, 658, A162
- Astropy Collaboration et al., 2013, *A&A*, 558, A33
- Astropy Collaboration et al., 2018, *AJ*, 156, 123
- Astropy Collaboration et al., 2022, *ApJ*, 935, 167
- Barragán O., Gandolfi D., Antoniciello G., 2019, *MNRAS*, 482, 1017
- Barragán O., Aigrain S., Rajpaul V. M., Zicher N., 2022a, *MNRAS*, 509, 866
- Barragán O. et al., 2022b, *MNRAS*, 514, 1606
- Batalha N. E., Lewis T., Fortney J. J., Batalha N. M., Kempton E., Lewis N. K., Line M. R., 2019, *ApJ*, 885, L25
- Beck J. G., 2000, *Sol. Phys.*, 191, 47
- Berdyugina S. V., 2005, *Living Rev. Sol. Phys.*, 2, 8
- Boisse I., Bonfils X., Santos N. C., 2012, *A&A*, 545, A109
- Bouchy F., Pepe F., Queloz D., 2001, *A&A*, 374, 733
- Bourrier V. et al., 2021, *A&A*, 654, A152
- Cegla H. M. et al., 2018, *ApJ*, 866, 55
- Chaplin W. J., Cegla H. M., Watson C. A., Davies G. R., Ball W. H., 2019, *AJ*, 157, 163
- Collier Cameron A. et al., 2019, *MNRAS*, 487, 1082
- Collier Cameron A. et al., 2021, *MNRAS*, 505, 1699
- Cosentino R. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84461V
- Crass J. et al., 2021, preprint ([arXiv:2107.14291](https://arxiv.org/abs/2107.14291))
- Cretignier M., 2022, PhD thesis, University of Geneva, Switzerland
- Cretignier M., Dumusque X., Allart R., Pepe F., Lovis C., 2020a, *A&A*, 633, A76
- Cretignier M., Francfort J., Dumusque X., Allart R., Pepe F., 2020b, *A&A*, 640, A42
- Cretignier M., Dumusque X., Hara N. C., Pepe F., 2021, *A&A*, 653, A43
- Cretignier M., Dumusque X., Pepe F., 2022, *A&A*, 659, A68
- Cretignier M., Dumusque X., Aigrain S., Pepe F., 2023, *A&A*, 678, A2
- Cretignier M., Pietrow A. G. M., Aigrain S., 2024, *MNRAS*, 527, 2940
- Delisle J. B., Unger N., Hara N. C., Ségransan D., 2022, *A&A*, 659, A182
- Desort M., Lagrange A. M., Galland F., Udry S., Mayor M., 2007, *A&A*, 473, 983
- Donati J. F., Brown S. F., 1997, *A&A*, 326, 1135
- Dumusque X., 2018, *A&A*, 620, A47
- Dumusque X., Udry S., Lovis C., Santos N. C., Monteiro M. J. P. F. G., 2011, *A&A*, 525, A140
- Dumusque X., Boisse I., Santos N. C., 2014, *ApJ*, 796, 132
- Dumusque X. et al., 2015, *ApJ*, 814, L21
- Dumusque X. et al., 2021, *A&A*, 648, A103
- Ervin T. et al., 2022, *AJ*, 163, 272
- Faria J. P. et al., 2022, *A&A*, 658, A115
- Fischer D. A. et al., 2016, *PASP*, 128, 066001
- Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
- Foukal P., 1998, *ApJ*, 500, 958
- Gelman A., Rubin D. B., 1992, *Statistical Science*, 7, 457
- Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2004, *Bayesian Data Analysis*, 2ndedn., Chapman and Hall/CRC
- Gibson S. R., Howard A. W., Marcy G. W., Edelstein J., Wishnow E. H., Poppett C. L., 2016, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 990870
- Gomes da Silva J., Santos N. C., Bonfils X., Delfosse X., Forveille T., Udry S., Dumusque X., Lovis C., 2012, *A&A*, 541, A9
- Hall R. D., Thompson S. J., Handley W., Queloz D., 2018, *MNRAS*, 479, 2968
- Hara N. C., Delisle J.-B., 2023, preprint ([arXiv:2304.08489](https://arxiv.org/abs/2304.08489))
- Hathaway D. H., 2010, *Living Rev. Sol. Phys.*, 7, 1
- Hatzes A. P., 1996, *PASP*, 108, 839
- Haywood R. D. et al., 2014, *MNRAS*, 443, 2517
- Haywood R. D. et al., 2022, *ApJ*, 935, 6
- Hogg D. W., Bovy J., Lang D., 2010, preprint ([arXiv:1008.4686](https://arxiv.org/abs/1008.4686))
- Howard R., Harvey J., 1970, *Sol. Phys.*, 12, 23
- John A. A., Collier Cameron A., Wilson T. G., 2022, *MNRAS*, 515, 3975
- John A. A. et al., 2023, *MNRAS*, 525, 1687
- Jones D. E., Stenning D. C., Ford E. B., Wolpert R. L., Loredó T. J., Gilbertson C., Dumusque X., 2017, preprint ([arXiv:1711.01318](https://arxiv.org/abs/1711.01318))
- Jurgenson C., Fischer D., McCracken T., Sawyer D., Szymkowiak A., Davis A., Muller G., Santoro F., 2016, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 99086T
- Klein B., Donati J. F., 2019, *MNRAS*, 488, 5114
- Klein B. et al., 2022, *MNRAS*, 512, 5067
- Klein B. et al., 2024, *MNRAS*, 527, 544
- Kochukhov O., 2016, in Rozelot J.-P., Neiner C., eds, Lecture Notes in Physics, Vol. 914, Cartography of the Sun and the Stars. Springer, Switzerland, p. 177,
- Lakeland B. S. et al., 2024, *MNRAS*, 527, 7681
- Leighton R. B., 1964, *ApJ*, 140, 1547
- Lienhard F., Mortier A., Buchhave L., Collier Cameron A., López-Morales M., Sozzetti A., Watson C. A., Cosentino R., 2022, *MNRAS*, 513, 5328
- Lienhard F., Mortier A., Cegla H. M., Cameron A. C., Klein B., Watson C. A., 2023, *MNRAS*, 522, 5862
- Lovis C. et al., 2011, preprint ([arXiv:1107.5325](https://arxiv.org/abs/1107.5325))
- Luger R., Bedell M., Foreman-Mackey D., Crossfield I. J. M., Zhao L. L., Hogg D. W., 2021, preprint ([arXiv:2110.06271](https://arxiv.org/abs/2110.06271))
- Malavolta L., Lovis C., Pepe F., Sneden C., Udry S., 2017, *MNRAS*, 469, 3965
- Mayor M., Queloz D., 1995, *Nature*, 378, 355

- Meunier N., 2021, preprint (arXiv:2104.06072)  
 Meunier N., Lagrange A. M., 2019, *A&A*, 625, L6  
 Meunier N., Lagrange A. M., 2020, *A&A*, 638, A54  
 Meunier N., Desort M., Lagrange A. M., 2010, *A&A*, 512, A39  
 Meunier N., Lagrange A. M., Borgniet S., Rieutord M., 2015, *A&A*, 583, A118  
 Meunier N., Pous R., Sulis S., Mary D., Lagrange A. M., 2023, *A&A*, 676, A82  
 Milbourne T. W. et al., 2019, *ApJ*, 874, 107  
 Mordasini C., Alibert Y., Georgy C., Dittkrist K. M., Klahr H., Henning T., 2012, *A&A*, 547, A112  
 Nicholson B. A., Aigrain S., 2022, *MNRAS*, 515, 5251  
 Noyes R. W., Hartmann L. W., Baliunas S. L., Duncan D. K., Vaughan A. H., 1984, *ApJ*, 279, 763  
 O'Sullivan N. K., Aigrain S., 2024, *MNRAS*, preprint (arXiv:2404.11662)  
 Pepe F. et al., 2021, *A&A*, 645, A96  
 Pesnell W. D., Thompson B. J., Chamberlin P. C., 2012, *Sol. Phys.*, 275, 3  
 Phillips D. F. et al., 2016, in Navarro R., Burge J. H., eds, Proc. SPIE Conf. Ser. Vol. 9912, Advances in Optical and Mechanical Technologies for Telescopes and Instrumentation II. SPIE, Bellingham, p. 99126Z,  
 Queloz D. et al., 2001, *A&A*, 379, 279  
 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. Vol. 2, MIT Press, Cambridge, Massachusetts  
 Rincon F., Rieutord M., 2018, *Living Rev. Sol. Phys.*, 15, 6

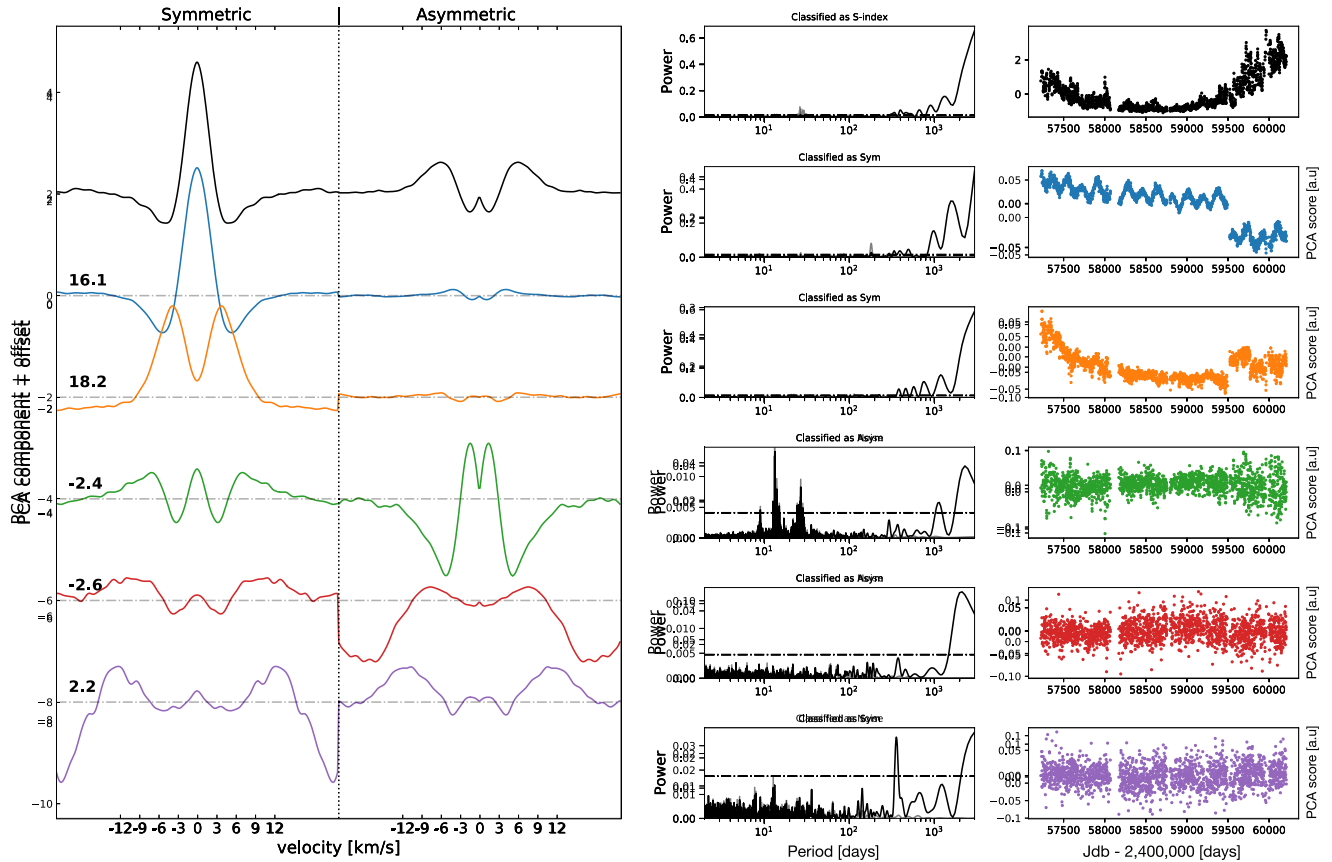
- Saar S. H., Donahue R. A., 1997, *ApJ*, 485, 319  
 Schwab C. et al., 2018, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII. SPIE, Bellingham, p. 1070271  
 Snodgrass H. B., Ulrich R. K., 1990, *ApJ*, 351, 309  
 Stalport M. et al., 2023, *A&A*, 678, A90  
 Thompson S. J. et al., 2016, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, 99086F  
 Thompson A. P. G. et al., 2020, *MNRAS*, 494, 4279  
 Wang Y. M., Nash A. G., Sheeley N. R. J., 1989, *ApJ*, 347, 529  
 Wilson T. G. et al., 2022, *MNRAS*, 511, 1043  
 Zechmeister M., Kürster M., 2009, *A&A*, 496, 577  
 Zhao L. L. et al., 2022, *AJ*, 163, 171  
 de Beurs Z. L. et al., 2022, *AJ*, 164, 49  
 de Beurs Z. L. et al., 2024, *MNRAS*, 529, 1047

## APPENDIX A: YARARA UPGRADE

The YARARA pipeline was initially described in Cretignier et al. (2021) as developed for the HARPS spectrograph. Its adaptation for HARPS-N was straightforward given that both instruments are very similar, but peculiarity of the HARPS-N instrument pushed us to slightly modify one of the post-processing recipe. Indeed, the YVI pipeline was subsequently improved to better disentangle the change of PSF from stellar activity (see appendix C in Stalport



**Figure A1.** PCA decomposition of the transformed CCFs residuals time-series. The figure is identical to the fig. C.2 obtained for HD4628 in Stalport et al. (2023) excepted that instead of a CCF obtained with all the stellar lines, we represented here the 'blue CCF' ( $\lambda < 4877 \text{ \AA}$ ) analysis. Left: PCA components obtained for the CCF residuals (see the main text). The component decorrelate by the  $S_{HK}$  is also shown in black. The vertical dotted line split the symmetric (left) from antisymmetric (right) profiles deformations. The parameter of symmetry  $R$  is shown on the left of the profiles and is in bold for significant symmetric profile ( $R > 3$ ). Right: Score of the PCA components and corresponding periodograms in days. The time-series coefficient used to correct for the change of the PSF are given by the symmetric components deformation (blue and orange components).



**Figure A2.** Same as Fig. A1 for the green CCF ( $4877 \text{ \AA} < \lambda < 5856 \text{ \AA}$ ). The jump corresponding for the warm-up of the detector is not visible around BJD = 58 629 on contrary to the blue CCF. This indicates that warm-up signatures are mainly visible in the blue part, coherent with the higher contrast of the ghosts.

et al. 2023). We briefly summarize the method used to measure the variation of the PSF time-series. We assumed that instrumental changes were mainly dominated by symmetrical variations of the line profile, which is valid since any asymmetric change of the line profile would also introduce an RV offset that is not observed. The method therefore consists in (1) deriving CCFs from the spectra, (2) correcting the position of their centroid by the RV value obtained from a Gaussian fit, (3) subtracting the median CCF from each RV-corrected CCF, (4) linearly detrending the median-subtracted CCFs from the  $S_{HK}$  index<sup>9</sup>, and (5) transforming the residuals CCFs  $\Delta\text{CCF}(v_i, t)$  into a merged symmetric versus asymmetric  $\Delta$  profiles  $\Delta = [\Delta\text{CCF}_{\text{sym}}(v_i, t), \Delta\text{CCF}_{\text{asym}}(v_i, t)]$ . A PCA is then performed on the transformed residual CCFs and only the components that are significantly symmetric (with a parameter<sup>10</sup>  $R > 3$ ) were selected. Such a method can be understood as an alternative version of the SCALPELS algorithm (Collier Cameron et al. 2021) using ACFs. The reason for the usage of the  $S_{HK}$  was the almost perfect relation between this index and the filling factor of active regions (Cretignier et al. 2024) and the relative unsensitivity of these lines to instrumental systematics that mainly affect sharps photospheric lines, but less strongly deep and broad chromospheric lines.

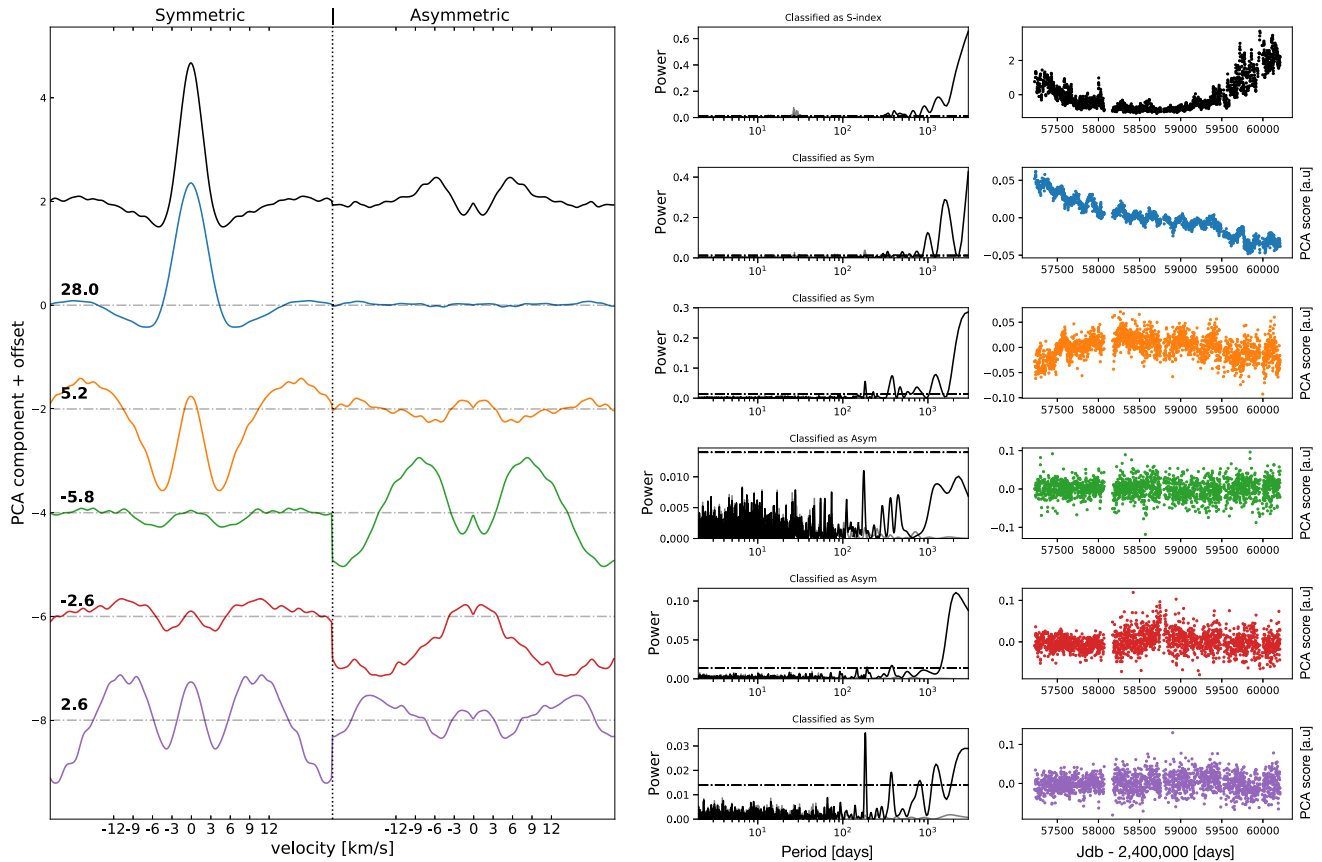
The time-domain scores of those components are then used as linear predictors to decorrelate the residual spectra time-series from

the instrumental PSF change as explained in Cretignier et al. (2021) where we already demonstrated that the method was preserving planetary signals. When we upgraded the recipe in Stalport et al. (2023), the change of the cryostat was at this time really recent and we did not have enough perspective to understand its effect, but the solar data revealed new properties about the instrumental intervention.

The previous recipe was mainly motivated to correct the large instrumental PSF defocus visible in 2012 at the beginning of the instrument’s lifetime (before the installation of the solar telescope). After the change of the cryostat, we observed that the previous developed recipe was imperfect. Indeed, the change of the PSF was previously obtained with a white CCF using the full wavelength bandpass between 3900 and 6835  $\text{\AA}$ , since at first order, all the stellar lines behave in a similar way on the spectrum. However, investigations on the Sun show that some instrumental effects are also very chromatic. As an example, the warm-up of the detector mainly affect the blue part of the spectrum since the contrast of the ghosts is larger in the blue spectral range. Also, the change of the cryostat introduced a larger signal in the green and blue part of the spectrum compared to the red, but the reason for this signature is yet unknown. We therefore modified the recipe developed in Stalport et al. (2023) by computing three colours CCFs (blue, green, and red) splitted evenly in wavelength and used the obtained vector to decorrelate the spectra residuals time-series as explained in YARARA Cretignier et al. (2021). The reason for three colour is mainly justified by the usual trade-off sensitivity versus SNR. We displayed the same figure as the one obtained for HD4628 in Stalport et al. (2023) for

<sup>9</sup>Where the linear coefficients are determined on the high-pass filter of the signals.

<sup>10</sup>see Stalport et al. (2023) for the precise definition of R



**Figure A3.** Same as Fig. A1 for the red CCF ( $\lambda > 5856 \text{ \AA}$ ). The jump corresponding for the warm-up of the detector is not visible, neither the cryostat change.

the solar observations in Figs A1, A2 and A3 with blue and green CCFs, respectively.

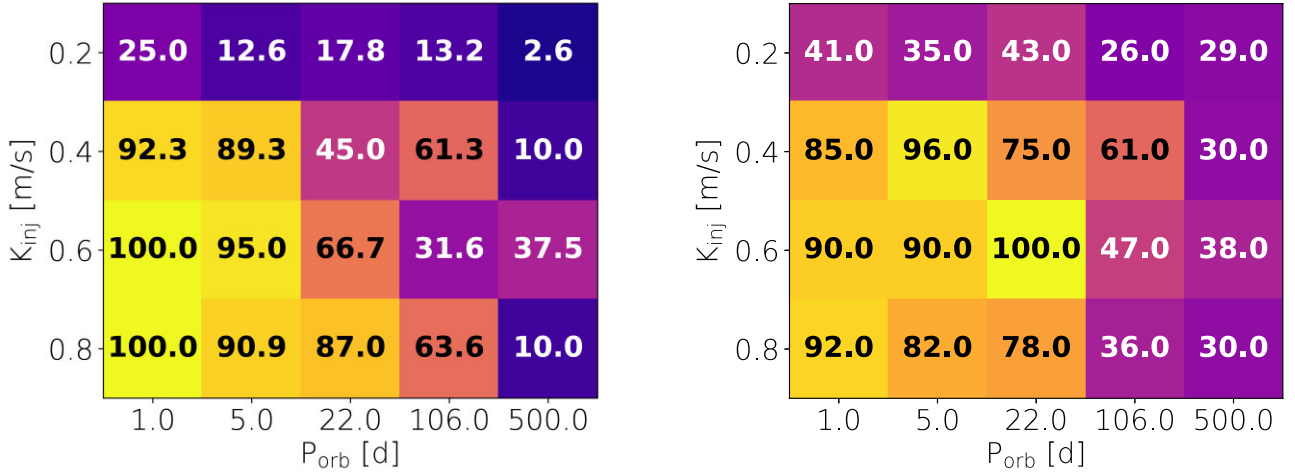
The first PCA component of blue and green CCFs reveal the clear signature of the change of  $v \sin i$  and of the cryostat change around  $\text{BJD} = 2459491$ . The second PCA component in the blue contains extra component such as the warm-up at  $\text{BJD} = 2458629$ , while the green CCFs contains a long trend with the cryostat offset. This long trend may be correlated with the long decrease in SNR due to ageing of the solar dome. The third PCA component is mainly an antisymmetric variation (therefore not used) and clearly show the stellar activity power at  $\text{Prot}/2$ .

## APPENDIX B: PLANET DETECTION MAPS

Fig. B1 shows the detection rates of the planets injected in the HARPS-N solar CCFs without and with activity filtering (see Section

4.2), as a function of the planet RV semi-amplitude  $K_{\text{inj}}$  and orbital period  $P_{\text{orb}}$ . Planets considered to be detected if the recovered RV semi-amplitude differs by less than  $1\sigma$  from  $K_{\text{inj}}$  and differs from 0 by at least  $3\sigma$ . As outlined in Section 4.2, the activity-filtering framework increases the sensitivity to planets with low semi-amplitude (typically lower than  $0.4 \text{ m s}^{-1}$ ) and larger orbital period (typically greater than  $\sim 100 \text{ d}$ ). We note a very light decrease of the detection rate of short-period higher-amplitude planet signatures after activity filtering. This can be understood as uncertainties on the RV semi-amplitude are significantly smaller when estimated from the activity-filtered RVs than from the raw RVs (see Fig. 9). Note that, the best estimates of the RV semi-amplitudes of the planets missed within this  $(K_{\text{inj}}, P_{\text{orb}})$  space still lie within  $2\sigma$  of the injected RV semi-amplitude.



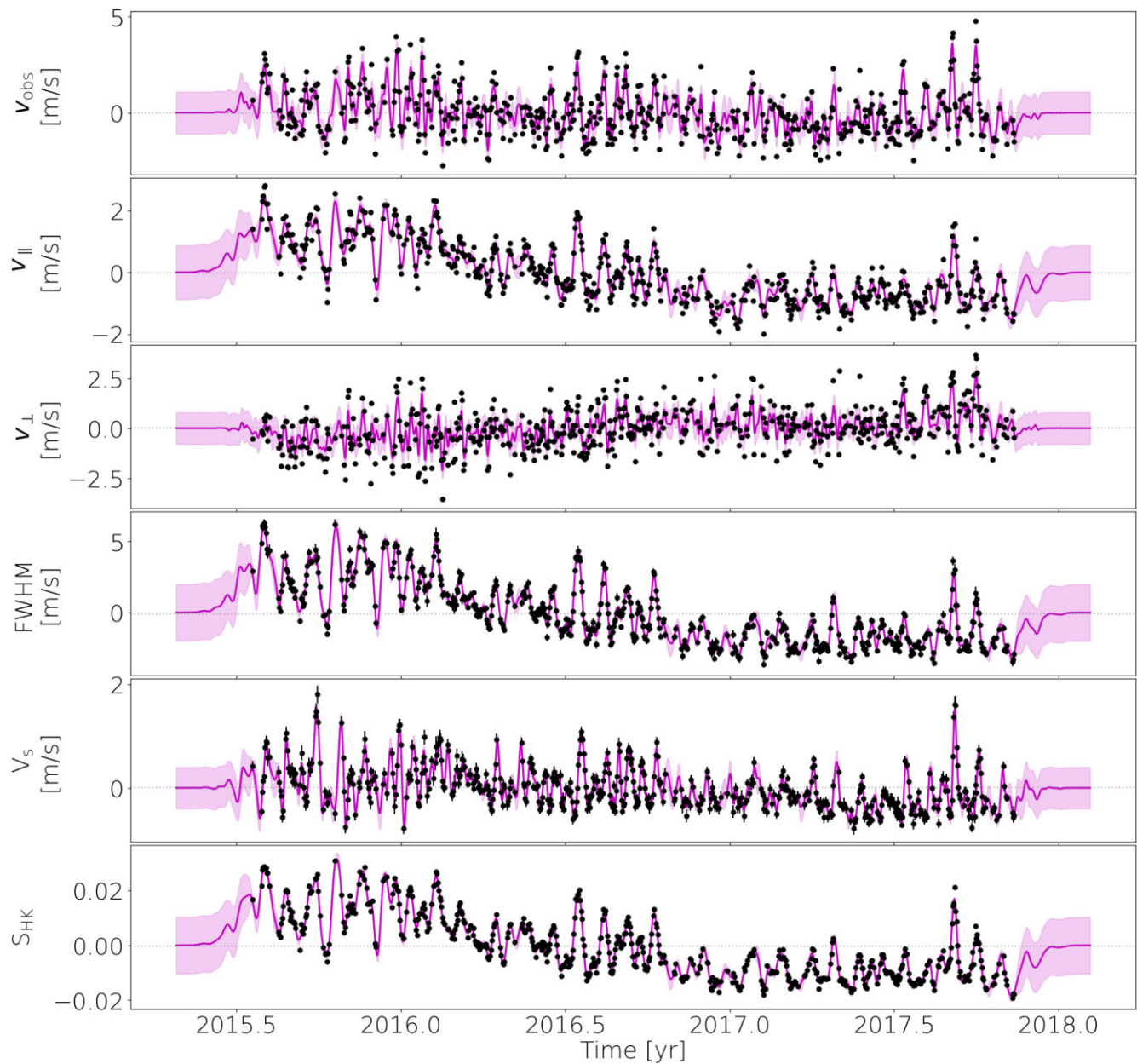


**Figure B1.** Planet detectability maps, as a function of the RV semi-amplitude and orbital period, recovered from the HARPS-N solar RVs without activity filtering (left-hand panel) and using the activity-filtering framework of Section 2.1. In each panel, the colour code depicts the planet detection rate (dark blue/yellow indicating detection rates close to 0/100 per cent). The numbers within each cell indicate the detection rate in per cent.

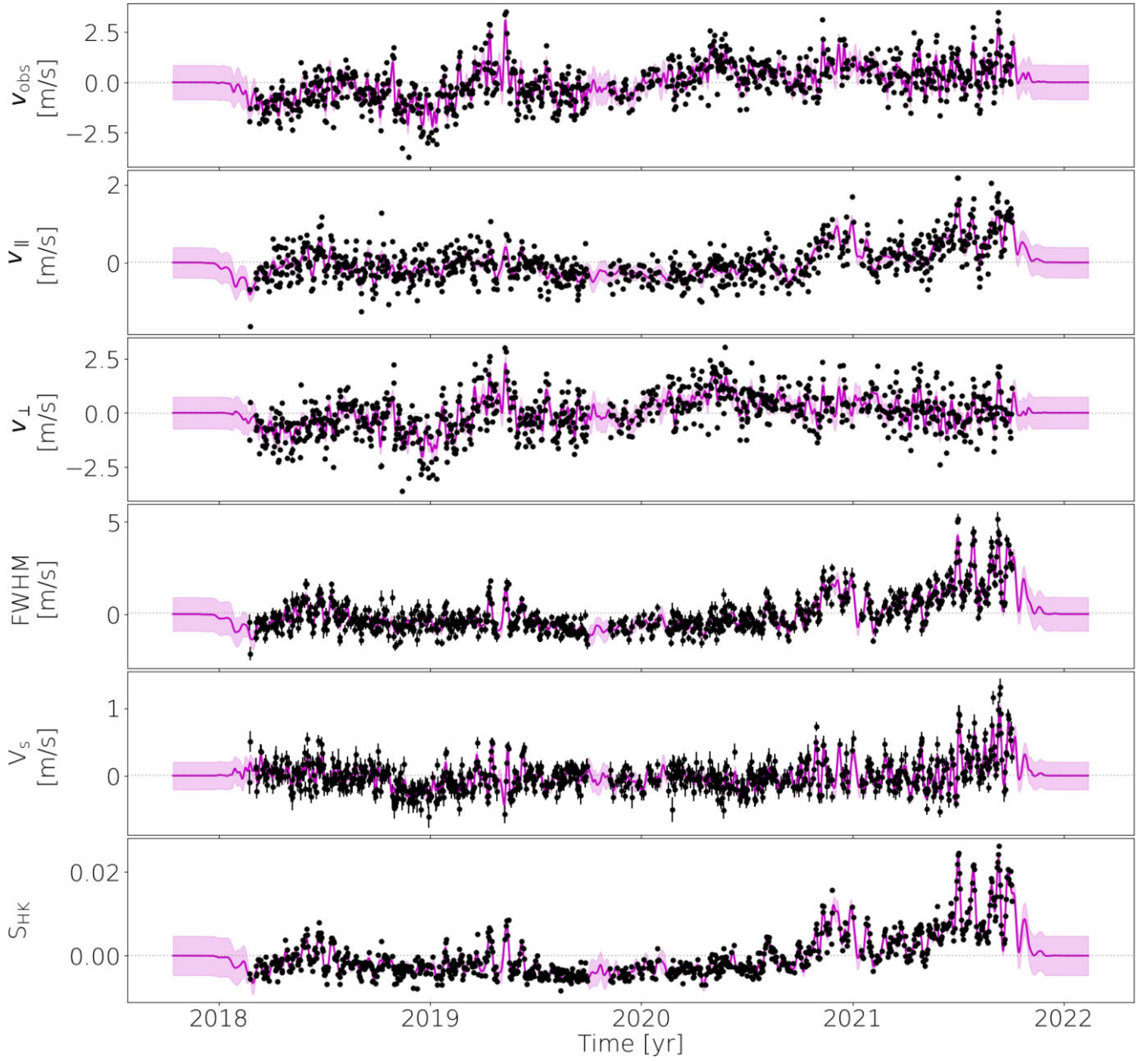
### APPENDIX C: GAUSSIAN PROCESS MODELLING

In this appendix, we give additional details of the GP modelling of spectroscopic activity indicators, namely the HARPS-N solar RVs ( $\mathbf{v}_{obs}$ ), the time-series of  $\mathbf{v}_{\perp}$  and  $\mathbf{v}_{\parallel}$ , computed in Section 4.1, and usual activity proxies (i.e. FWHM,  $V_s$  and  $S_{HK}$ ). In Figs C1, C2, and C3, we show the quasi-periodic 1D GP fit to the different time series in Season 1 (2015–2018), 2 (2018–2021.8), and 3 (2021.8–2024), respectively. The posterior densities of the hyper-parameters of the 1D GP fit to the time-series of eigenvectors  $\mathbf{U}_1$ ,  $\mathbf{U}_3$ , and  $\mathbf{U}_4$ , extracted in Section 4.1, are shown for all three seasons in Fig. C4. We visually see the change in the solar activity regime from one season to the next. Activity-induced fluctuations are most clearly in the start of cycle 25 (Season 3, Fig. C3), when the Sun is more

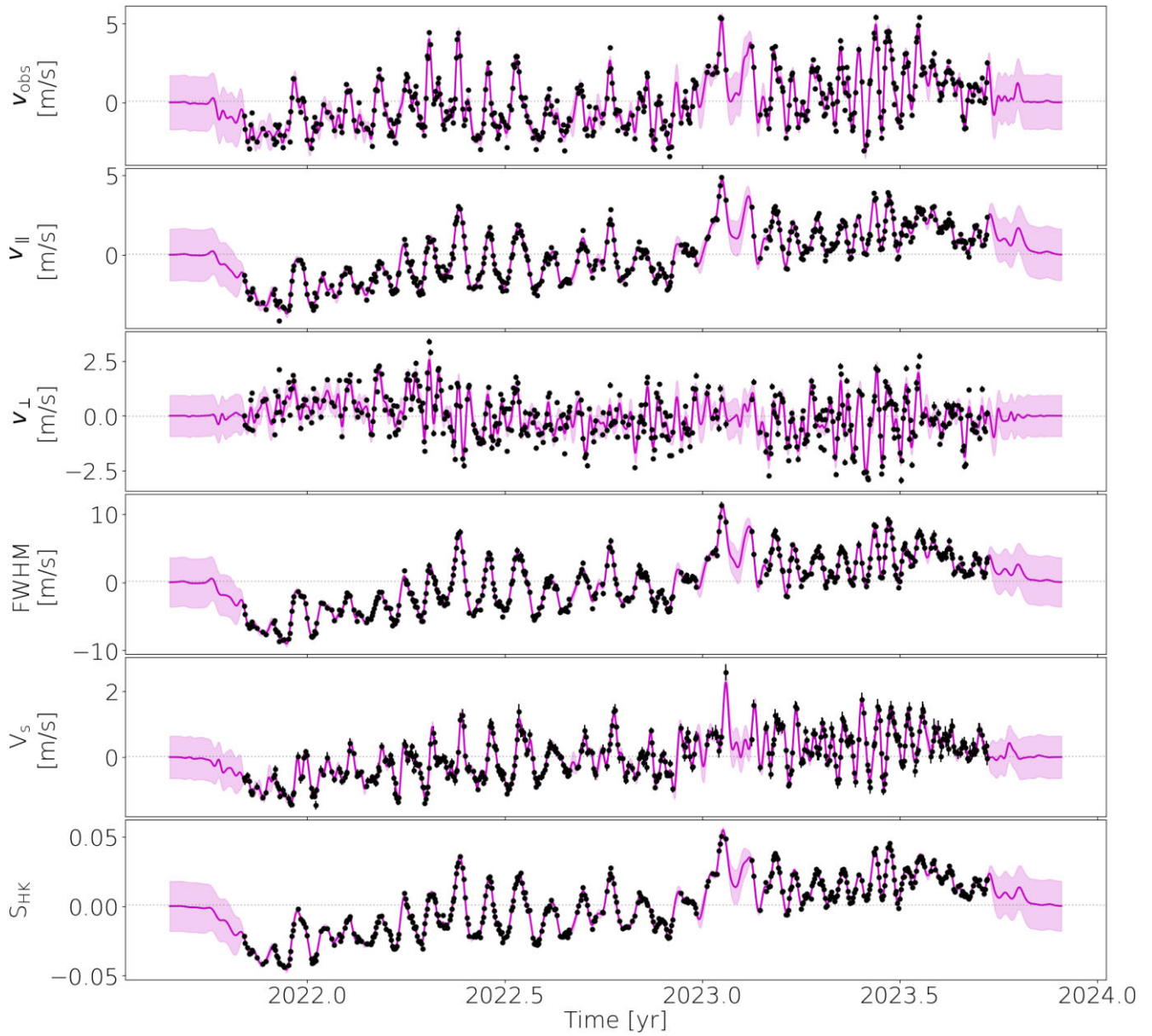
active. In this regime, rotationally modulated signals are present in all activity proxies except  $\mathbf{v}_{\perp}$ , where they have been filtered out. These activity-induced fluctuations appear significantly faster-evolving at the end of cycle 24 (Season 1, Fig. C1), which is consistent with the lower GP evolution time scale and inverse harmonic complexity of this season (i.e.  $\lambda_e$  and  $\lambda_p$  from Table 4). The signal is the most complex over the solar minimum (Season 2, Fig. C2). In this case, the GP fit is controlled by the few activity-induced fluctuations, visible notably in the end of the season, which explains why the best-fitting GP parameters in Table 4 are similar to those obtained in Season 3. Outside these activity-dominated regions, we note that the FWHM and  $S_{HK}$  index are mostly flat, whereas the RVs still exhibit fluctuations of about  $1 \text{ m s}^{-1}$  peak-to-peak, most likely due to an interplay between activity residuals, granulation and instrument stability.



**Figure C1.** Best-fitting GP prediction to different time-series of activity indicators extracted from HARPS-N Solar spectra, during Season 1 (i.e. 2015 to 2018, see Table 2). In each panel, the data points with formal  $1\sigma$  error bars are shown in black dots, and the best-fitting GP prediction (resp.  $1\sigma$  error bands) is indicated by the magenta solid line (resp. shaded bands).

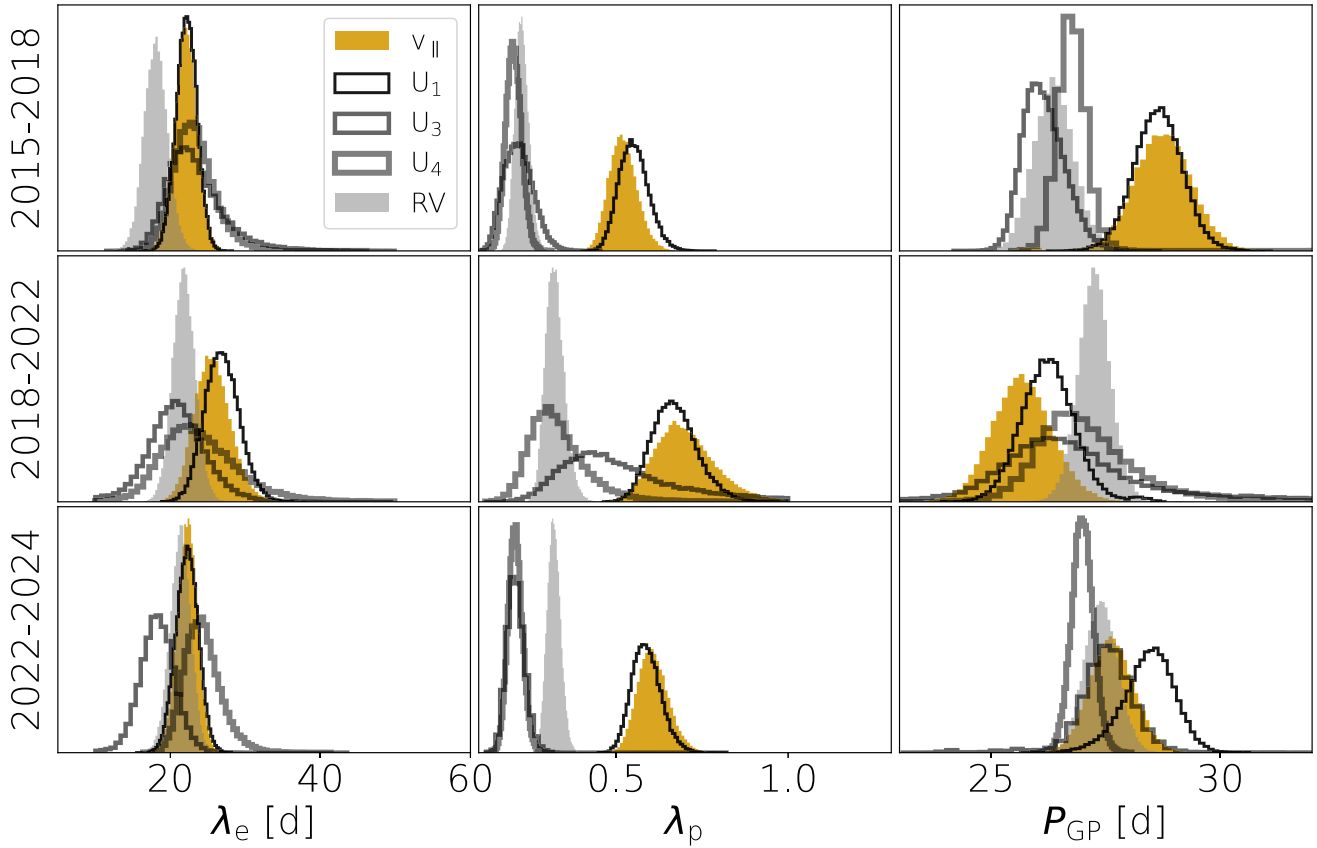


**Figure C2.** Same as Fig. C1, for the activity indicators observed during Season 2 (i.e. during solar minimum; see Table 2).

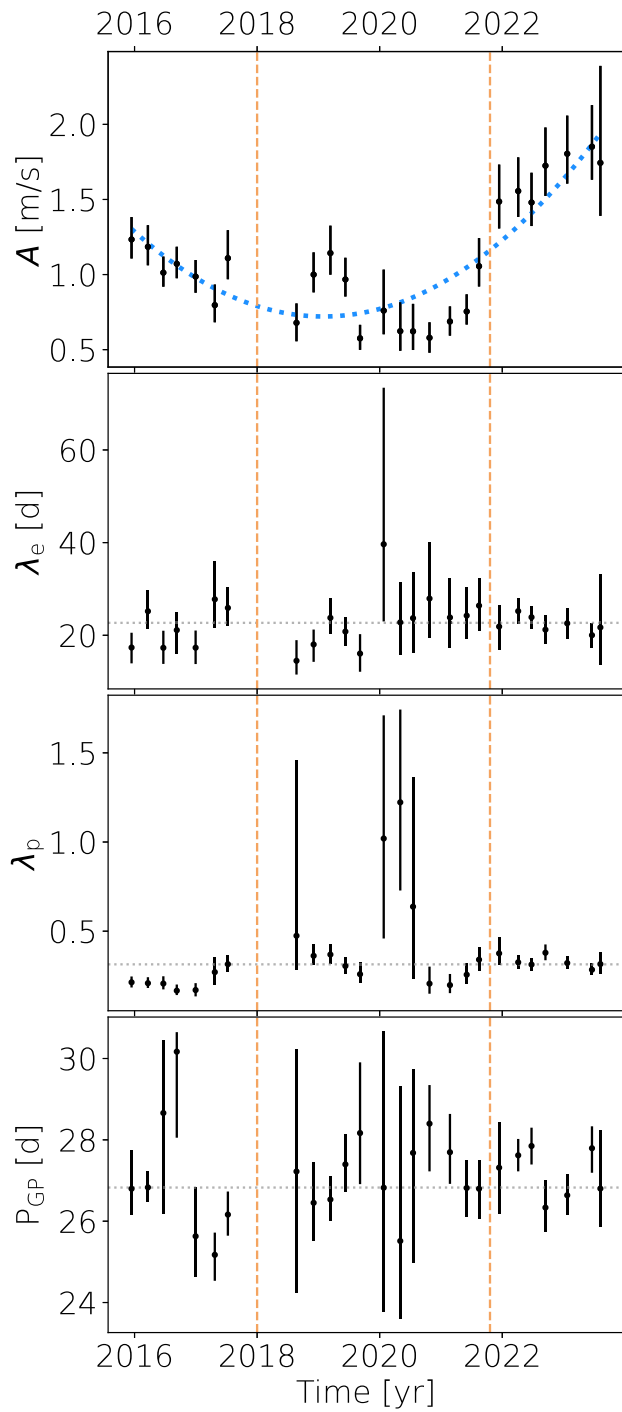


**Figure C3.** Same as Fig. C1, for the activity indicators observed during Season 3 (i.e. start of cycle 25; see Table 2).





**Figure C4.** Uni-dimensional posterior distributions of the GP evolution time-scale (left-hand column), inverse-harmonic complexity (middle column) and period (right-hand column) obtained by modelling the HARPS-N solar RVs (filled grey histograms),  $v_{\perp}$  (filled yellow histograms), and  $U_1$ ,  $U_3$ , and  $U_4$  (thin, medium thick, and thick grey lines).



**Figure C5.** Distribution of the best-fitting GP hyperparameters obtained by modelling 270-d chunks of HARPS-N solar RV. The dotted blue line in the top panel shows the best-fitting parabola to the GP amplitudes. The three seasons defined in Table 2 (i.e. end of Cycle 24, solar minimum and start of Cycle 25) are delimited by the vertical dashed lines.

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.