

Measuring masses of transiting Earth-like planets: first insights from the PLATO RV data challenge

M. Cretignier¹,¹★ N. C. Hara,² S. Aigrain,¹ S. Sulis,² F. Bouchy,³ S. T. Hodgkin,⁴ J. Blanco-Pozo,^{5,6,7} A. Collier-Cameron¹⁰,⁸ M. Damasso,⁹ E. C. Gonçalves,¹⁰ L. Malavolta¹⁰,^{11,12} G. Mantovan¹⁰,¹² D. Nardiello¹⁰,^{11,12} M. Perger¹⁰,^{5,6} I. Ribas,^{5,6} Y. Zhao,¹³ X. Dumusque,³ A. Mortier¹⁴, K. Al Moulla¹⁰,¹⁰ S. C. C. Barros,¹⁰ N. C. Santos,^{10,15} C. S. N. Martinez,¹⁰ C. Watson,¹⁶ T. Wilson¹⁷ and S. Udry³

the PLATO Stellar Variability Working Group

¹Department of Physics, University of Oxford, OX13RH Oxford, UK

²Aix Marseille Univ, CNRS, CNES, LAM, 13388 Marseille, France

³Astronomy Department of the University of Geneva, 51 ch. de Pegasi, CH-1290 Versoix, Switzerland

⁴Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK

⁵Institut de Ciències de l'Espai (ICE, CSIC), Campus UAB, Carrer de Can Magrans s/n, E-08193 Bellaterra, Barcelona, Spain

⁶Institut d'Estudis Espacials de Catalunya (IEEC), Edifici RDIT, Campus UPC, E-08860 Castelldefels, Barcelona, Spain

⁷Facultat de Física, Universitat de Barcelona (UB), Martí Franquès 1, E-08028, Barcelona, Spain

⁸SUPA School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK

⁹INAF – Osservatorio Astrofisico di Torino, Via Osservatorio 20, I-10025 Pino Torinese, Italy

¹⁰Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, P-4150-762 Porto, Portugal

¹¹Dipartimento di Fisica e Astronomia ‘Galileo Galilei’ – Università degli Studi di Padova, Vicolo dell'Osservatorio 3, I-35122, Padova, Italy

¹²Centro di Ateneo di Studi e Attività Spaziali ‘G. Colombo’ – Università degli Studi di Padova, Via Venezia 15, I-35131, Padova, Italy

¹³Department of Astronomy, University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA

¹⁴School of Physics & Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

¹⁵Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

¹⁶Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK

¹⁷Department of Physics, University of Warwick, Coventry CV4 7AL, UK

Accepted 2026 August 12. Received 2026 August 12; in original form 2026 May 27

ABSTRACT

The PLATO Stellar Variability Mitigation Working Group designed a radial velocity (RV) data challenge to assess the precision and accuracy achievable on RV semi-amplitudes in the context of the future PLATO follow-up. Using Sun-as-a-star observations obtained with HARPS-N, we generated three multiplanet systems with realistic sampling, an effective signal-to-noise ratio of ~ 450 , an exposure time of $T_{\text{exp}} = 15$ min, and a total of $N \sim 1000$ observations. The resulting data set is publicly available and can serve as a benchmark for the community to test and compare RV analysis methods. Eight independent teams participated in the challenge, aiming to recover the semi-amplitudes of both transiting and non-transiting planets. A meta-analysis of the submitted results shows that, with current data analysis methods, HARPS-N-like stability and the observing strategy studied in this paper, achieving a precision on semi-amplitudes better than 10 per cent for habitable-zone signals around a solar-type star is challenging, and is typically limited to planets with masses of ~ 8 – $10 M_{\oplus}$. This performance is expected to improve for K dwarfs, where planets down to ~ 2 – $5 M_{\oplus}$ may be characterized at a similar level of precision. Finally, based on the results of this challenge and the occurrence of false positives, we derive an empirical detection threshold for RV surveys, which can be used to distinguish robust planetary detections from candidates.

Key words: line: profiles – methods: data analysis – techniques: radial velocities – techniques: spectroscopic – stars: activity.

1 INTRODUCTION

The transit method is one of the most efficient and prolific techniques for detecting exoplanets today. Previous space missions such as *Kepler* (launched in 2009, W. J. Borucki et al. 2010; D. G. Koch et al. 2010) and *TESS* (launched in 2018, G. R. Ricker

* E-mail: michael.cretignier@physics.ox.ac.uk

et al. 2015) have considerably improved our understanding of planetary demographics, particularly for short-period planets (E. A. Petigura, A. W. Howard & G. W. Marcy 2013; N. M. Batalha 2014; C. J. Burke et al. 2015; S. Bryson et al. 2020; R. L. Royer & J. H. Steffen 2025; K. Cui et al. 2026).

The transiting configuration naturally biases detections towards short-period planets. However, thanks to their large fields of view monitoring thousands of stars simultaneously, both missions have produced numerous and prominent detections. This success is largely due to the fact that the signature of a transiting planet is highly orthogonal to most stellar photometric variability signals, such as granulation and rotational modulation. However, these two missions have shortcomings. Most stars in the *Kepler* field are faint, which makes the mass measurements difficult. Although *TESS* targets are brighter, the observation strategy focuses on short-period planets.

PLANetary Transits and Oscillations of stars (PLATO) is an upcoming space-based photometric mission designed to observe bright stars, such that follow-up observations are possible, and its observation strategy is designed to enable the detection of Earth analogues (H. Rauer et al. 2014, 2024, 2025). Indeed, the survey strategy is broadly similar to that of *Kepler*, with a long-duration pointing towards a southern field observed for at least two years. This baseline enables the detection of at least two transits of Earth-like planets, or potentially single-transit (mono-transit) events. The nominal mission is expected to continue for an additional two years, either by switching to a northern field or by extending the southern observations, resulting in a total time baseline comparable to *Kepler*'s four-year monitoring. Furthermore, its photometric precision is expected to be close to *Kepler*, with an anticipated noise level of about 20–30 ppm on hourly time-scales for bright targets (N. Jannsen et al. 2024). Unlike *Kepler*, however, PLATO will primarily target brighter stars (J. Cabrera et al. 2026; M. Montalto et al. 2026), a key advantage for enabling precise ground-based RV follow-up and mass characterization of the detected planets.

The radial velocity (RV) follow-up is an integral part of the mission and aims at measuring the mass of the exoplanets detected by PLATO, including Earth analogues. Combined with the radius estimate, this will enable a mean density measurement and yield constraints on the interior structure. Breaking the degeneracy between several interior compositions requires a high accuracy on the density, ideally ~ 10 per cent for terrestrial planets (M. Plotnykov & D. Valencia 2024), which translates to stringent requirements on the measurement of RV semi-amplitude, as seen below.

1.1 The PLATO mission requirements on mass measurements

The PLATO mission requirements R-SCI-Lo-15¹ reads: *PLATO shall observe stars that are bright enough to allow for a determination through radial velocity measurements of the mass of a terrestrial planet orbiting a G0V star with an accuracy of 10 per cent or better. ‘Terrestrial planet’ is defined as having $R < 2 R_{\oplus}$ and $M < 10 M_{\oplus}$ – a goal that remains challenging for current RV capabilities.*

In principle, the recent ultra-stable spectrographs are capable of reaching such a precision with sub m s^{-1} stability. These include ESPRESSO (F. Pepe, D. Ehrenreich & M. R. Meyer 2014),

NEID (C. Schwab et al. 2018), EXPRES (C. Jurgenson et al. 2016), HARPS-N (R. Cosentino et al. 2012), MAROON-X (A. Seifahrt et al. 2018), and KPF (S. R. Gibson et al. 2016), building upon the legacy of instruments such as HARPS (M. Mayor et al. 2003). For instance, ESPRESSO has demonstrated an intra-night precision of the order of 10 cm s^{-1} (F. Pepe et al. 2021; P. Figueira et al. 2025). Unfortunately, at these precision levels, stellar activity dominates the RV signal, significantly complicating the detection of Earth-like planets around Sun-like stars (see e.g. N. Meunier & A. M. Lagrange 2020a; J. Crass et al. 2021; N. C. Hara & E. B. Ford 2023).

The challenge in RVs is greater than in photometry because the planetary signal is not a so-called gated signal (i.e. a component that appears and disappears over time), but instead is continuously superposed with all other contributions. This has important implications: for instance, machine-learning algorithms cannot be trained on subsets of the data that are ‘free’ of the planetary signal, but also it means that any incomplete or inaccurate model will always interact with the planetary signals in unpredictable ways. In addition, the contaminating signals are more numerous, often larger in amplitude, and more problematic in the time and frequency domains (N. Meunier et al. 2015; D. A. Fischer et al. 2016; J. Crass et al. 2021; M. Cretignier et al. 2021; N. C. Hara & E. B. Ford 2023).

Some attempts have been made to evaluate the RV limitations based on simulations (N. Meunier & A.-M. Lagrange 2019; N. Meunier & A. M. Lagrange 2020a, b), but the results are subject to caution. On the one hand, the simulations might not be realistic enough; on the other hand, the data analysis methods used in these studies might be too simplistic. Far from being discouraging, they motivate further intensive investigations based on real data sets and gathering more independent experts: ultimately the organization of an RV challenge.

In this paper, we aim to assess whether the PLATO mass-precision requirements can be fulfilled with current methods and the extent to which they can be met across the (K, P) parameter space. We have therefore designed an RV challenge involving several competitors and collaborators within the PLATO community that uses some state-of-the-art methodologies currently available. This challenge aims to mimic as closely as possible a PLATO follow-up. One of its most important aspects is that it makes use of real data, where planets are injected at the spectral level.

In Section 2, we introduce the main concept of the challenge. We begin by reviewing previous challenges (Section 2.1), followed by a description of the design of the present challenge, including the injected planetary systems and the data sets used (Section 2.2). The participating teams and their methods are then presented in Section 2.3. The results are described in Section 3, including the definition of the scoring metrics (Section 3.1), the performance on the different data sets (Section 3.2), and the overall ranking of the methods (Section 3.3). We then discuss the broader implications for the RV field (Section 3.4), before concluding in Section 4.

2 A PLATO RV CHALLENGE

2.1 Previous RV challenges

Few RV challenges have been conducted over the past decade. The main advantage of such community exercises is that they allow different methods to be tested on identical data sets. This

¹<https://platomission.com/2018/05/19/top-scientific-requirements>

enables a meaningful benchmarking of the field, provides a snapshot of the current state of the art, and allows the performance of the various approaches to be assessed within a consistent framework.

The first RV challenge was organized by X. Dumusque et al. (2017) using simulations as well as real HARPS data sets in which planetary signals were injected directly into the $RV(t)$ time-series. Although realistic, the simulations were not capturing all the intricacies of stellar variability. Indeed, a linear fit of activity indicators worked extremely well to remove stellar activity signals in simulated systems, but performed worse in practice. Furthermore, this challenge was unfortunately limited to high-level data products from the data reduction software, namely the extracted RV time-series and common activity proxies. In recent years, however, several studies have suggested that significant improvements can be achieved by working directly in the velocity domain (Z. L. de Beurs et al. 2020; J. Zhao & C. G. Tinney 2020; A. Collier Cameron et al. 2021a; M. Cretignier, X. Dumusque & F. Pepe 2022; J. Zhao, E. B. Ford & C. G. Tinney 2022; H. Yu et al. 2024; B. Klein et al. 2025), in the wavelength domain (M. Cretignier et al. 2021; S. Shahaf & B. Zackay 2023; L. L. Zhao et al. 2024) or by combining multiple RV time-series derived from different spectral line selections (F. Lienhard et al. 2022; K. Al Moulla, X. Dumusque & M. Cretignier 2024; D. Dravins & H.-G. Ludwig 2024) or obtained with different extraction methods (M. Zechmeister et al. 2018; A. M. Silva et al. 2022).

More recently, L. L. Zhao et al. (2022) organized another challenge based on four EXPRES data sets, comparing different methodologies. In this case, data products at multiple processing levels were made available to participants, offering a broader and more up-to-date perspective on the field. However, the conclusions were less definitive. First, the metric used to compare methods is the root mean square of the corrected RVs, which might favour methods that are overfitting (N. C. Hara & E. B. Ford 2023). Second, no participant consistently outperformed the others across all four data sets, which complicates any attempt at establishing a robust ranking metric.

Although it is well known that the true planetary content of real data sets can never be fully known since low undetected signals may always exist,² the absence of injected signals in any of the data sets represents a limitation of the analysis. Injecting realistic planetary signals at the low-level data stage is technically challenging – and propagating those signals consistently through to high-level products is even more complex. However, we are aware that such an attempt is in progress (Zhao et al., in preparation).

2.2 Design of the PLATO challenge

We designed the PLATO RV challenge to closely mimic the conditions expected for the RV follow-up of PLATO targets, implementing some recommendations of N. C. Hara & E. B. Ford (2023) on the design of RV challenges. To achieve this, several high-level design choices were made.

First, the challenge exclusively relies on real data sets. While simulated data allow complete control over the ground truth, they inevitably lack certain complexities present in real observations, such as instrumental systematics and subtle correlated

noise sources. These effects remain part of the real problem and should not be underestimated.

Second, planetary signals corresponding to transiting planets are injected at the low-level data stage on the spectra, rather than directly into derived RV time-series. This ensures that injected signals propagate consistently through the full reduction and analysis chain.

Third, to place the RV challenge within PLATO's core objective of detecting Earth-mass planets in the habitable zone (HZ) of Sun-like stars, we ensured that at least one transiting exoplanet is always located in the Sun's HZ. This guarantees that the target star would be considered a high-priority candidate for follow-up within the Ground-based Observation Program of PLATO.

Finally, additional planetary signals from non-transiting planets are injected. This enables the methods to be evaluated under realistic conditions, where both known (transiting) and unknown (non-transiting) signals coexist. As a reminder, in the Solar System no more than three planets can be seen transiting from any given vantage point in the Universe (R. Wells et al. 2018). This is important as the number of measurements needed to estimate the mass at a given precision of transiting exoplanets increases if non-transiting companions are present (M. Y. He, E. B. Ford & D. Ragozzine 2021). Also, the detection of non-transiting planets is a more stringent test of RV methods. While it seems possible for a data analysis method to retrieve by chance a small semi-amplitude, close to that of an injected planet, the odds of retrieving by chance the correct period of the injected signal are very low.

2.2.1 Injected planetary system

The parameter space of multi-Keplerian systems is extremely large. In this challenge, we therefore restricted ourselves to the production of three multiplanetary data sets and limited the analysis to purely circular orbits. Without loss of generality, the term Keplerian will be used hereafter, even if it should be understood as a *sinusoid*, rather than a strict Keplerian. All the participants were aware of the null eccentricity hypothesis and therefore removed it from their Keplerian fit adjustment. This is certainly one of the least realistic aspects of the challenge, since in this configuration only the semi-amplitudes of the transiting signals need to be adjusted. However, as shown in Y. Shen & E. L. Turner (2008) and R. V. Baluev (2015), it is not significantly more difficult to detect eccentric planets than circular ones up to $e \approx 0.7$. This is also the limit from which computational schemes might struggle to estimate eccentricities (N. C. Hara et al. 2019). There is no clear evidence that such eccentricities are common for low-mass planets.

Transit measurements are known to yield far more precise constraints on phase T_0 and orbital period P than RV observations. In practice, this precision would differ from system to system, adding an additional layer of complexity to the challenge analysis – particularly given the limited number of data sets we generated. Therefore, for all the transiting planets, we provided the exact orbital periods and transit phases such that all the participants set their priors on them to an arbitrarily large precision. The performance of the teams will be solely evaluated based on their ability to retrieve the semi-amplitudes of the transiting planets, and find the transiting signals.

Based on Doppler shift injection tests performed in Appendix B, we restricted the values of the K semi-amplitudes below ~ 25

²Exceptions due to stability consideration may exist in very few specific cases (see e.g. L. L. Zhao et al. 2021).

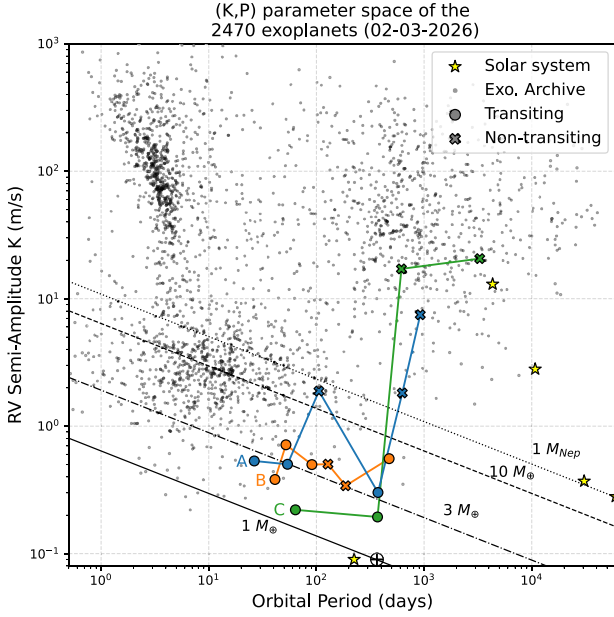


Figure 1. Illustration in the (K, P) parameter space of the three multiplanetary systems injected in the HARPS-N solar data: A (blue), B (orange), and C (green). Transiting planets are depicted with a round marker (see the legend). All the transiting planets in this work were smaller than $7M_{\oplus}$. Four isomasses lines are indicated. The 2470 exoplanets with a measured semi-amplitude according to the NASA Exoplanet Archive are represented as a comparison.

m s^{-1} for the injected planets. Above this threshold, the Doppler wavelength shift interpolation used to inject the planets in the spectra may lose their accurate validity to a degree that is critical for the RV challenge.

The three injected multiplanetary systems are represented in the (K, P) parameter space in Fig. 1 and are hereafter referred to as systems A, B, and C. The Keplerian parameters are indicated in Table 1. All terrestrial planets range between 2 and $7M_{\oplus}$, which are already challenging planets to characterize today.

For all the systems, we always assumed the solar properties for the host star, including the HZ limits even if we extrapolated the results to cooler type stars later (see Section 3.5).

2.2.1.1 System A

System A is a hierarchical six-planet system composed of three transiting terrestrial planets, two non-transiting Neptune-sized planets, and one non-transiting Saturn-mass planet.

The system architecture is inspired by the HIP 41 378 system, which hosts both transiting and non-transiting planets spanning a wide range of orbital periods, and thus provides a realistic benchmark for complex planetary configurations (e.g. A. Vanderburg et al. 2016; A. Santerne et al. 2019; S. Grouffal et al. 2025). To adapt this architecture to our study, we rescaled the orbital periods by a factor of 1.7 for all planets, except for the fourth planet.

As a result, the orbital period of the innermost planet is close to the solar rotation period around 27 d, providing a stringent test for disentangling planetary signals from stellar activity due

Table 1. Summary table of the three multiplanetary systems A, B, and C injected in the PLATO RV challenge. Information available to the participants is indicated in bold. The orbital period P , RV semi-amplitude K , planetary mass m_p and transit phase T_c are indicated. The type column indicates to which category a planet belongs (T = transit, M = mono-transit). Note that both systems A and C have their HZ signal close to 365 d, which implies an incomplete phase coverage due to season gaps. The semi-amplitude K_{raw} obtained by fitting the system on the data without the planet injected is also indicated with the relative amplitude with injected signals.

Syst.	P	K	m_p	T_c	Type	K_{raw}
	(d)	(m s^{-1})	($M_{\oplus}$)	(BJD)		(m s^{-1}) (percent)
A	26.47	0.53	2.49	57152.2818	T	0.11 (21)
A	53.90	0.50	2.97	57163.1609	T	0.15 (30)
A	105.5	1.88	13.9	57166.2604	–	0.11 (6)
A	373.0	0.30	3.4	57142.0194	T	0.12 (40)
A	627.3	1.82	24.4	57186.9142	–	0.27 (15)
A	921.5	7.48	114	57000.0000	–	0.92 (12)
B	41.21	0.38	2.06	54971.5052	T	0.05 (13)
B	52.10	0.71	4.17	54971.1748	T	0.11 (15)
B	90.74	0.50	3.51	54981.4550	T	0.20 (40)
B	128.0	0.50	3.97	54987.1590	–	0.20 (40)
B	186.8	0.34	3.05	54964.6487	–	0.15 (44)
B	473.5	0.56	6.78	55120.2901	M	0.29 (52)
C	63.79	0.22	1.38	59640.1579	T	0.13 (59)
C	369.5	0.19	2.18	59646.0919	T	0.35 (184)
C	615.9	17.1	228	59641.5763	–	0.80 (5)
C	3286	20.6	480	59656.0921	–	2.86 (14)

to increased signal crosstalk in both photometric and RV time-series.

The third planet is non-transiting, similarly to HIP 41 378 g, and contributes only through dynamical and RV signatures. The outermost planet is also non-transiting and induces a long-period dominant signal in the RV data.

The fourth planet was adjusted to transit and to lie within the HZ of a solar-type (G2V) star, corresponding to orbital periods between 338 and 586 d. To satisfy this constraint while preserving the system hierarchy, a specific scaling factor of 1.36 was applied to its orbital period. This planet is designed to mimic a high-priority PLATO candidate suitable for follow-up observations.

Planetary radii were scaled by factors between 2 and 2.5 in order to control the resulting RV semi-amplitudes. A relation $m_p \propto R_p^{3.7}$ was assumed to convert radii to masses following (L. Zeng, D. D. Sasselov & S. B. Jacobsen 2016). By doing so, we ensured that all injected signals remain below 1 m s^{-1} , consistent with the low-amplitude regime expected for PLATO targets.

2.2.1.2 System B

System B is a compact six-planet system composed exclusively of rocky planets, inspired by the *Kepler-11* system (J. J. Lissauer et al. 2011). Such systems are representative of tightly packed inner planetary systems commonly observed by transit surveys, making them a relevant benchmark for testing detection performance in crowded configurations.

The system includes four transiting planets (the three innermost and the outermost) and two non-transiting planets. The outermost planet is located within the HZ. In this configuration, the habitable-zone planet produces a mono-transit event. Only

the transit epoch was provided, while the orbital period was left unconstrained,³ mimicking realistic observational limitations.

The planetary radii were scaled by factors of 1.6, 2.1, 2.4, 2.5, 1.9, and 2.1, while all orbital periods were increased by a factor of 4. This makes all the RV semi-amplitudes below 70 cm s^{-1} , placing the system in a regime where stellar activity dominates the RV signal, thereby significantly complicating planet detection and characterization.

2.2.1.3 System C

System C is a four-planet system composed of two transiting terrestrial planets and two long-period non-transiting giant planets.

The system architecture was inspired by the HD 110 067 system (R. Luque et al. 2023), with modified orbital periods and two planets removed to simplify the configuration (planets d, e, f, g only). Their orbital periods were rescaled by factors of 7, 12, 15, and 60. Planetary radii were kept unchanged, except for the innermost planet, whose radius was reduced by a factor of 2.

The two inner planets are terrestrial and transiting. The second planet is located within the HZ to mimic a high-priority PLATO target for follow-up observations. The two outer planets are long-period, non-transiting giant planets.

Despite the reduced number of injected planets compared to Systems A and B, the terrestrial signals in System C are among the most challenging to detect, with RV semi-amplitudes as low as 20 cm s^{-1} .

The outermost giant planet has an orbital period close to the 11-yr solar magnetic cycle (Schwabe cycle), introducing additional difficulty in disentangling long-period planetary signals from stellar magnetic variability.

2.2.2 HARPS-N solar data set

To create the RV challenge, we used the HARPS-N solar telescope data set (X. Dumusque et al. 2015, 2021). HARPS-N has been observing the Sun daily since 2015 with 5-min integration times continuously during the day. We used the data release corresponding to X. Dumusque et al. (2021), although a more recent release became available this year (X. Dumusque et al. 2026). The baseline of the observations spans eight years from July 2015 (BJD = 2457220) up to October 2023 (BJD = 2460238).

The use of HARPS-N solar data offers several advantages compared to archival stellar RV data bases. A key benefit is that, for the Sun, the planetary signals of the Solar System can be removed, thereby providing a well-defined ground truth. In addition, the exceptional temporal sampling of the solar observations enables flexible control of the cadence, allowing it to be reshaped in a realistic manner to closely match the expected sampling of the PLATO RV follow-up.

However, several important caveats must be considered when extrapolating solar results to stellar observations.

First, although difficult to quantify, it should be noted that some participants may have previously used or worked with the HARPS-N solar data set. This potentially affects the strictly ‘blind’ nature of the challenge,⁴ and may therefore introduce a slight

imbalance between participants. This limitation is arguably an inherent trade-off when relying on a data set of such importance, which has been widely used in the literature over the past years.

The data sets used in the challenge have an average visible bandpass⁵ signal-to-noise ratio (S/N) of 450 ± 40 . The impact of a lower S/N on the results was not explored, although such an analysis would be valuable and could be performed through the injection of photon noise.

Third, because the Earth orbits the Sun, Sun-as-a-star data sets are less sensitive to some instrumental systematics that are fixed in the laboratory rest frame. When observing the Sun, the barycentric Earth radial velocity (BERV) span is limited to approximately $\sim 2 \text{ km s}^{-1}$ (due to the Earth’s orbital eccentricity), whereas for typical stellar observations the BERV span can reach dozens of km s^{-1} . Given that the typical photospheric line width of a slow-rotating star is $\sim 6 \text{ km s}^{-1}$ for an HARPS-like instrumental resolution, the instrumental systematics in solar observations remain effectively ‘frozen’ with respect to the line profile, whereas in stellar observations they are swept across the spectrum over time. This difference alters how instrumental effects couple to the spectral lines (see e.g. A. M. Silva et al. 2025).

Last but not least, the solar telescope follows a different optical path than standard stellar observations. On the one hand, this removes some sources of systematics related to telescope guiding (see e.g. the recently unperfected fibre scrambling identified on HARPS-N in A. Anna John et al. 2025). On the other hand, the disc-resolved nature of the Sun introduces additional complications, particularly differential extinction across the solar disc. Although this effect has been calibrated at the RV level through an analytical correction term $\text{RV}_{\text{corr}}(t)$ in A. Collier Cameron et al. (2019), its correction at the spectral level has never been fully demonstrated or implemented. As a result, a simple Doppler shift by the measured $\text{RV}(t)$ is not strictly sufficient to correct the spectra.

Aware of this issue, we limited this effect doing the following selection of the spectra. First of all, we only selected spectra having a good quality flag. Bad data are usually identified as outliers in the daily magnitude versus airmass relation which is expected to be linear (K. Al Moulla et al. 2023) and as such diagnose the periods in time when clouds occulted the Sun. We restricted the effect of the differential extinction by only selecting solar observation close to the meridian with an airmass smaller than $Z = 1.3$. At such an airmass, the differential extinction effect is lower than 10 cm s^{-1} according to the $\text{RV}_{\text{corr}}(t)$ vector. Note that this choice also recreates a realistic season-length sampling since the Sun in winter is not going below an airmass $Z = 1.3$.

Each day, we randomly selected three consecutive observations that satisfy the previous requirements to create an effective 15-min exposure time, which mimics the expected exposure time for PLATO targets and strongly mitigates the contribution from the 5-min p-mode pulsations (W. J. Chaplin et al. 2019). Stacking the three consecutive spectra produces an effective visual bandpass S/N ~ 450 or 15 cm s^{-1} of RV photonic precision. As an important note, the Sun-as-a-star on HARPS-N⁶ is equivalent to a magni-

signals between data sets may hint at injected planets. This is in particular the case for the giant ones.

⁵On HARPS-N, the average S/N of the visual bandpass equals 83 per cent of the one at $5500 \text{ \AA}$.

⁶Explained by the smaller telescope size and the use of the integration sphere.

³A weak prior was given, indicating that the period was expected to lie between 200 and 500 d.

⁴The use of the same data set across multiple realizations also partially compromises the blindness of the challenge, as the absence or presence of

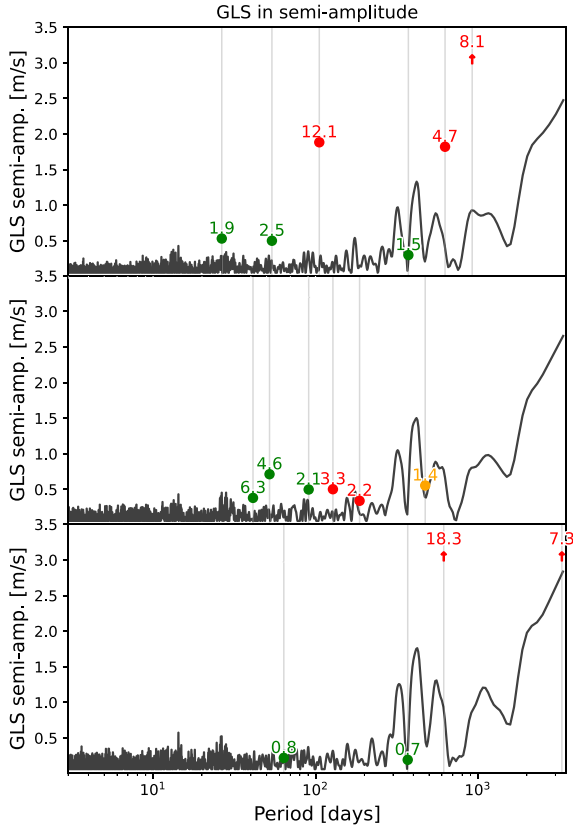


Figure 2. Illustration of the GLS periodogram for the three data sets [A1050 (top panel), B1300 (middle panel), and C800 (bottom panel)] without the planets injected. Although the GLS is traditionally represented in terms of power, we display it here in K semi-amplitude sinusoids scaling to facilitate a direct comparison with the semi-amplitudes of the injected signals. Transiting planets are shown in green (mono-transit in orange), and non-transiting ones in red. Signals larger than 3 m s^{-1} are just truncated and their marker is replaced by an upper arrow. We also indicate above each marker the ratio of the planetary amplitude to the native GLS power at the same period.

tude $m_v = 5.4$ on the 3.6m TNG telescope or to a magnitude $m_v = 7.2$ on the 8.2m VLT telescope. Such a magnitude represents the bright tail of the PLATO sample (V. Nascimbeni et al. 2026).

In order to avoid obvious reverse engineering (such as taking the difference of two RV time-series), three independent realizations of spectra selection were done for each of the planetary systems described in Section 2.2.1. Each data set contains a different number of observations. We will call those three data sets (hereafter A1050, B1300, and C800) related to the respective planetary system and the number of data points each of those data sets has (see Fig. 3 for the exact number). Note that the data sets A, B, and C represent a huge time investment with 26, 33, and 21 equivalent nights of telescope time, respectively, which is realistically the amount of time needed to be able to measure such HZ signals. The season length S of each data set is also slightly different even though longer than half-a-year, with $S = 225$, $S = 236$ and $S = 207 \text{ d}$. This is roughly two months shorter than the season length expected from PLATO field from Paranal ($S \sim 280 \text{ d}$).

We displayed in Fig. 2 the generalized Lomb–Scargle (GLS) periodogram (M. Zechmeister & M. Kürster 2009) in semi-amplitude of the three RV data sets without the planets injected and also indicate the ratio between the planetary semi-amplitude

and the native semi-amplitude power at this period. Similarly, we fitted the planetary systems on the RVs without the planet injected in order to quantify the initial amount of power and degeneracy between the injected systems and the data sets. We reported the semi-amplitudes K_{raw} in Table 1: the larger the number, the higher the cross-talk between the native power spectrum and planetary signals.

The RV time-series of the three data sets with the planets injected are shown in Fig. 3. As a comparison, we also displayed, in the panel below, the MgII index⁷ converted in plage filling factor using the time-series published in M. Cretignier et al. (2024a). We also displayed the solar total solar irradiance (TSI; see Section 2.2.8) that is a mixture of faculae and sunspot filling factor.

2.2.3 Processing of the data

Injection of the planetary signals at the spectrum level may be challenging since, to be exact, the planetary signal should be injected on spectra free of instrumental systematics. Indeed, an incorrect processing that would strongly underestimate the EPRV problem would be to simply shift the spectra by the Keplerian solution, injecting the planetary signals on the laboratory rest-frame systematics.

We used YARARA (M. Cretignier et al. 2021) that is a post-processing pipeline analysing spectra time-series to correct instrumental systematics at the spectrum level. As a short recall, YARARA starts with a preliminary standardization of the SID order-merged spectra by fitting their continuum with the RASSINE code (https://github.com/MichaelCretignier/Rassine_public; M. Cretignier et al. 2020b). During this stage, all the spectra are also interpolated linearly on a common wavelength grid in order to analyse and correct each wavelength bin’s time-series independently. The code then creates a master spectrum and fits the residual spectra time-series multilinear bases from proxies extracted by the pipeline itself.

The YARARA correction of HARPS-N includes corrections of cosmics rays, telluric lines, ghosts (X. Dumusque et al. 2015), interference pattern (e.g. A. P. G. Thompson et al. 2020), instrumental point spread function (PSF) distortions (see M. Stalport et al. 2023 or B. Klein et al. 2024) such as those introduced by the change of the solar $v \sin i$ along the year (A. Collier Cameron et al. 2019) and stellar activity. Note that any correction can be cancelled and added back on the spectra easily. This is mandatory since the stellar activity and change of the PSF should be on the spectra prior to their RV shift by the Keplerian solution.

For each data set, we first ran YARARA on the spectra time-series, removed the stellar activity and PSF correction, shifted the spectra according to the Keplerian solutions, before adding back all the contaminations. We then ran a second time the pipeline on this new data set and offer to the participant two flavours of the data: (a) the YV0 flavour that is uncorrected of any systematics and (b) the YVA flavour corrected of all the instrumental systematics except the stellar activity. Participants were free to choose between the two versions of the data they wanted to use. The data ($\sim 15 \text{ GB}$) are freely accessible on Google Drive⁸ and Zenodo.⁹

⁷https://lasp.colorado.edu/lisird/data/bremen_composite_mgii

⁸<https://drive.google.com/drive/u/1/folders/1Jbz-i21D811lr90w3AVSDzygctGuFnqo>

⁹<https://doi.org/10.5281/zenodo.20151946>

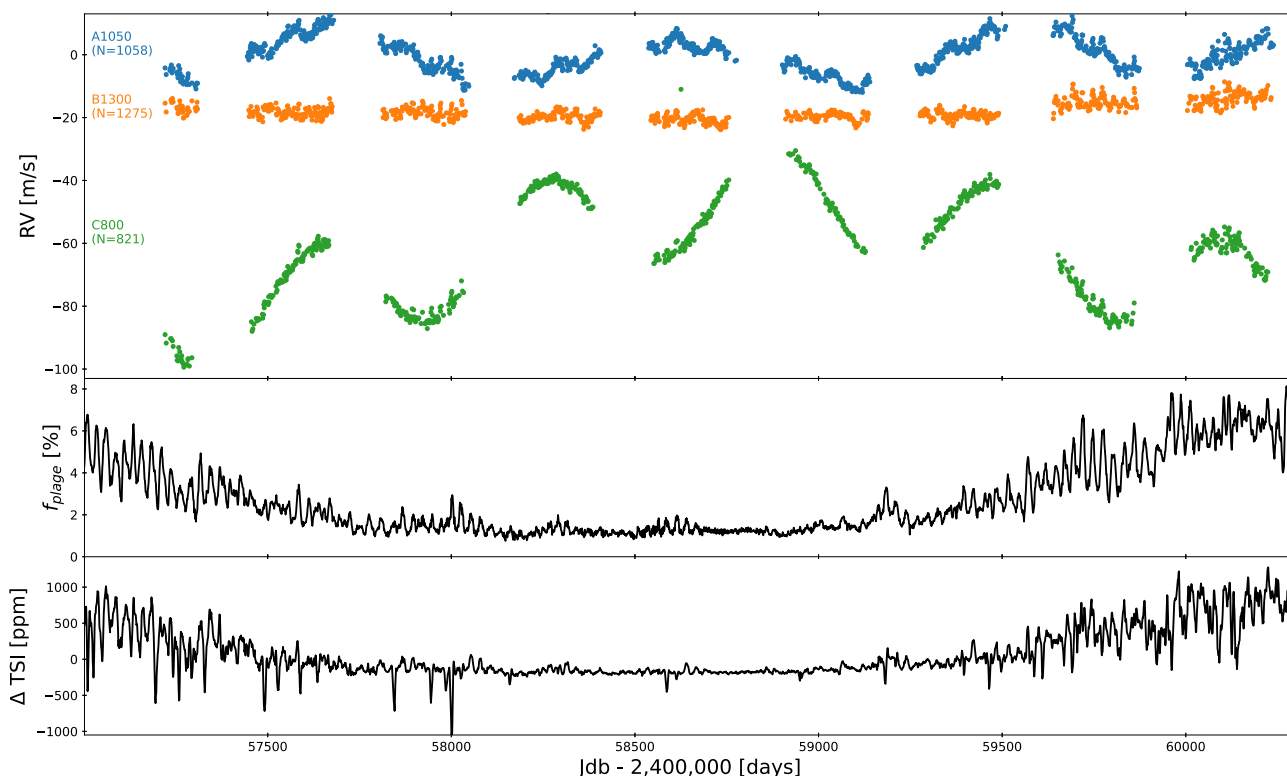


Figure 3. Representation of the main time-series provided for the RV challenge. Top panel: RV time-series (YV0) of the A1300 (blue), B1050 (orange), and C800 (green) data sets. Arbitrary vertical offsets were applied for graphical considerations. Middle panel: The plage filling factor as extracted from the MgII index and converted here into a plage filling factor using the reference time-series of M. Cretignier, A. G. M. Pietrow & S. Aigrain (2024a). Bottom panel: Total solar irradiance converted here from usual W/m^2 units to ppm by using solar minimum flux level as reference.

2.2.4 HARPS-N solar S-index

We provide the S-index time-series extracted from the HARPS-N spectra, which is a standard diagnostic typically delivered alongside RVs for visible-band spectrographs. The S-index is directly related to the filling factor of active regions (M. Cretignier et al. 2024a), although the line response remains relatively weak for solar-like activity, requiring careful spectral extraction.

Indeed, the Sun-as-a-star observations obtained with HARPS-N are known to be affected by significant contamination from optical ghosts crossing the cores of the Ca II H&K lines, from which this activity proxy is derived (X. Dumusque et al. 2021).

For the Sun, this contamination would in principle be of limited concern, as instrumental systematics are expected to remain largely stable over time. However, regular warm-up cycles of the HARPS-N detector modify the relative intensities of these ghost reflections with the main physical orders, introducing not only an absolute contamination, but also a time-variable component.

This contamination can be mitigated by estimating the flux between physical spectral orders on the CCD, as described in X. Dumusque et al. (2021), or by using the YARARA framework, which includes dedicated corrections for such effects (M. Cretignier et al. 2021). We opted for the latter for simplicity.

In addition, the contamination is further reduced by applying the emission-profile extraction method presented in M. Cretignier et al. (2024b). In contrast to the classical Mount Wilson S-index, which relies on a triangular bandpass flux integration, this approach directly fits on the spectra the emission profile of plage regions to the data (see B. J. Oranje 1983; M. Cretignier et al. 2024a). This method provides improved uncertainties and is less

sensitive to outliers compared to standard bandpass integration. In this framework, the plage emission profile used for the extraction can be interpreted as an optimally weighted integration of the signal.

Note that, due to the spectrograph spectral efficiency response of HARPS-N, the S/N at 550 nm is equivalent to twice the value of the one in the continuum region surrounding the CaII H&K lines at 420 nm.

2.2.5 Solar TSI

We provided the TSI as obtained from the LISIRD¹⁰ data base. We did not inject the different transits in the TSI because the latter is not necessarily representative of the precision expected from the PLATO light curves, adding an extra layer of complexity in the challenge and its possible interpretation that we wanted to avoid. Instead of that, as explained previously, we provided the exact phase and period of the transiting systems and mainly provided the TSI as a proxy that participants could use to fit their stellar activity model.

2.2.6 ARVE solar photospheric components

In order to benchmark the RV challenge and assess performance on a less demanding data set, we generated a simplified red-noise model free of instrumental systematics and active regions: it only

¹⁰https://lasp.colorado.edu/lisird/data/nnl_tsi_P1D

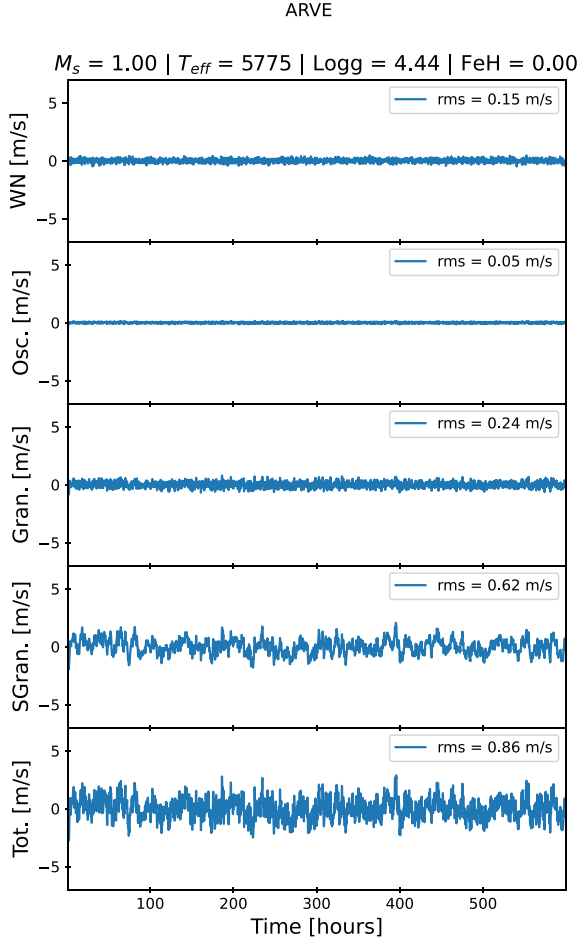


Figure 4. Representation of the RV budget expected from the main solar components in the PLATO RV challenge that uses the solar HARPS-N data set. A time-series of eight years including 15 min of integration time was generated using the ARVE code with the solar atmospheric parameters (only one rotational period is shown here). The contribution from each component is indicated in the legend from top to bottom: the photon noise (15 cm s^{-1}), p-mode oscillation (5 cm s^{-1}), granulation (24 cm s^{-1}), supergranulation (62 cm s^{-1}), and their sum (86 cm s^{-1}).

contains white noise, oscillations, granulation and supergranulation. For this purpose, we used the ARVE code (<https://github.com/almoulla/arve>; K. Al Moulla 2025), which simulates realistic time-series from their power spectrum, as fitted to the HARPS-N solar data set. The individual RV contributions are shown in Fig. 4 for a three-week baseline. The model predicts a minimum total RV jitter of 86 cm s^{-1} , composed of 15 cm s^{-1} from photon noise, 5 cm s^{-1} from p-modes, 24 cm s^{-1} from granulation, and 62 cm s^{-1} from supergranulation. These contributions combine to produce the total RV budget of 86 cm s^{-1} similar to the budget found in K. Al Moulla et al. (2023) and N. Meunier & A. M. Lagrange (2020b). We note that we did not include potential long-term magnetic cycle modulation of these components. Recent work suggests that supergranulation properties may vary with magnetic activity (N. K. O’Sullivan et al. 2025), but this effect is not accounted for in the present model. Using ARVE, we simulated the same eight-year RV time-series with 15-min integration times as the three data sets. This theoretical red-noise time-series will be used (see Section 2.3.1) for benchmarking purposes and represents an optimistic case scenario.

2.3 Participants

In total, eight participants participated in the challenge: Seven provided a full set of results and one participant provided a partial result analysing only one (B1300) of the three systems. We compared the results of the candidates with a fictive processing represented by a benchmark participant. The output expected from the participants were the K_i posterior samples from their inferred Keplerian solution. We present here the methodologies of all the participants to deliver those products.

2.3.1 Benchmark (exact Keplerian solution + ARVE)

This fictive participant is not using the YV0 or YVA data sets, but the ARVE RV budget described in Section 2.2.5 that includes the same planetary signals on it. A Keplerian signal was added for each transiting planet with the period and phase specified by the challenge. Non-transiting planets were added if visible as a significant peak in the residuals periodograms (obtained after fitting the transiting planets’ amplitudes), which was the case for all the signals due to the lack of power for the ARVE time-series at long ($P > 3 \text{ d}$) periods. All the eccentricities were assumed null, and a uniform prior was used on all the remaining parameters. Note that we also investigated a benchmark with free eccentricities that provided very similar results, and we therefore stuck to the null eccentricity model for a fair comparison with the participants.

Under the assumption that no participant is able to deal with supergranulation (which is a question none of them have tried to answer so far), this participant represents the best performance that can be achieved in theory since the Keplerian model is close to the exact one and there is no long-term stellar activity and instrumental systematics. It will allow us to check if the amount of information in the time-series was sufficient or if solar photospheric components are sufficient to hinder the success of the challenge. All the Markov chain Monte Carlo (MCMC) analyses were done using the DACE¹¹-implemented tool. Transit phases and periods were kept fixed for the transiting planets during the MCMC exploration.

2.3.2 Linear regression of the S-index (S. Aigrain)

The participant analysed both the YV0 and YVA data sets, assuming that all planets were on circular orbits, as the participants were informed of this a priori. The analysis consisted of two steps: (1) detecting the non-transiting planets and (2) jointly fitting for the planetary and stellar signals.

In both steps, the stellar RV variations were assumed to be proportional to the variations of the S-index. In step 1, the RVs were first modelled as a linear combination of the S-index time-series, and sinusoidal terms corresponding to the known ephemerides of the transiting planets (ignoring mono-transits). The coefficients of the resulting linear basis model were fitted using SCIKIT-LEARN.LINEARREGRESSION¹² (version 1.2.2), and the best-fitting model was subtracted from the RVs. The GLS periodogram of the residuals was then computed using ASTROPY.TIMESERIES.LOMBSCARGLE¹³ (version 6.1.7), on a regularly spaced grid of frequencies from $\nu = 0.0002$ to $0.9 \text{ cycles d}^{-1}$.

¹¹<https://dace.unige.ch/radialVelocities/>

¹²<https://scikit-learn.org>.

¹³<https://www.astropy.org/>.

($P = 1.1\text{--}5000$ d), with `nyquist_factor` = 10. If the highest peak in the periodogram had a False Alarm Probability $FAP \leq 10^{-5}$, the corresponding period was taken to correspond to an additional planet. A sine and a cosine term at this new period were added to the linear basis model, and the process was repeated until no further signals were found below the FAP threshold.

The fitted coefficients of these additional terms were then converted to RV semi-amplitudes and T_0 values for the additional planets. For system 3, whose RV variations are dominated by non-transiting planets rather than activity signals or transiting planets, the initial fit was omitted and the initial GLS periodogram was computed directly on the YV0/YVA RVs.

Step 2 consisted of a Bayesian modelling of the planetary and stellar signals using Hamiltonian Monte Carlo (HMC) to explore the joint posterior over all the parameters. The free parameters of the fit were thus: the systemic RV_0 offset, the semi-amplitudes K_i of all the planets, the periods P_i and times of inferior conjunction $T_{0,i}$ of the additional planets identified at step 1 only, the coefficient S of the S -index term, and a jitter term σ_{jit} added in quadrature to the RV uncertainties. This participant used JAXOPLANET¹⁴ (version 0.1.0) to evaluate the RV signatures of the planets.

The HMC sampling was implemented in NUMPYRO¹⁵ (version 0.18.0) using the No U-Turn Sampler. For the semi-amplitudes, half-normal priors with a mean of zero and a standard deviation of 50 m s^{-1} were used. For the periods, the prior was a truncated log-normal with a standard deviation of 0.05 (in log), and bounds mid-way between the logarithms of the initial period guesses. This choice of prior allowed the model to explore a fairly wide range of periods for each planet without allowing the planets to ‘switch’ order. A uniform prior between 0 and 2π was used for the orbital phase of the additional planets, using NUMPYRO-EXT.DISTX.ANGLE¹⁶ (version 0.0.4) to sample this circular variable correctly. Wide normal priors (with a standard deviation of 10 m s^{-1}) were used over RV_0 and S , and a log-uniform prior between 0.1 and 3 m s^{-1} over σ_{jit} . After a 1000-step burn-in phase, two chains of 10 000 steps each were run, which was sufficient to ensure that the Gelman-Rubin statistic $R \leq 1.01$ and the effective sample size was more than half of the number of samples for all parameters and all three systems. The resulting samples were provided as output along with various diagnostic plots (best-fitting model and residuals, chains and corner plots).

2.3.3 Linear regression of the S -index (M. Damasso)

The team analysed the YVA data set only, assuming that all planets have circular orbits, which is a prior information provided to the participants. For all the targets, each test model included a linear regression with the S_{MW} index and the planetary Doppler signals, after verifying that a significant correlation exists between the RV and S_{MW} data sets. They also included a constant uncorrelated jitter term, which was added in quadrature to the nominal RV uncertainties.

A frequency content analysis of the RV data was initially performed with GLS to identify potential signals due to non-transiting planets in the original data sets and pre-whitened residuals, having a $FAP < 0.01$ per cent determined through a bootstrap analysis. These signals defined the maximum number of planets

to be fitted for each simulated system through a Monte Carlo (MC) simulation. The MC modelling was performed by exploring the parameter space with the publicly available nested sampler and the Bayesian inference tool MULTINEST v3.10 (e.g. F. Feroz et al. 2019), through the PYMULTINEST wrapper (J. Buchner et al. 2014), included 500 live points, a sampling efficiency of 0.3, and a Bayesian evidence tolerance of 0.5. The package RADVEL (v1.4.9; B. J. Fulton et al. 2018) was used to model the RV Doppler planetary signals.

Bayesian model comparison was performed by calculating the difference $\Delta \ln \mathcal{Z}$ between the values of the Bayesian evidence $\ln \mathcal{Z}$ given by MultiNest for each model, assuming that the a priori probability is the same for each model, and following the scale in F. Feroz, S. T. Balan & M. P. Hobson (2011) to interpret the statistical significance $\Delta \ln \mathcal{Z}$. Values of $\Delta \ln \mathcal{Z}$ were calculated between pairs of models with a number of planets n and $n + 1$. The team adopted uninformative priors for all the free parameters, except for the orbital period and epoch of inferior conjunction of the transiting planets, for which they adopted Gaussian prior distributions using the information provided by the organisers. In particular, we explored the semi-amplitudes of the planetary Doppler signals within intervals with a lower limit of zero and upper limits from 5 to 50 m s^{-1} , depending on the χ -square best-fitting values calculated with GLS for each potential planetary signal.

The team analysed the residual signal in RVs and found a significant stellar activity component that could be fit with a quasi-periodic GP. However, inclusion of the GP itself with the previously fit Keplerian systems did not significantly affect their results and a decision was made ultimately to adopt in this work models with no GP term included in order to provide results from a framework not too sophisticated that can be regarded as a benchmark against which to compare those obtained using more complex models.

2.3.4 Multi-GPs (L. Malavolta, D. Nadiello, G. Mantovan)

The participant analysed the YVA data set only and performed a joint modelling of photometric and spectroscopic data within a Bayesian framework using PyORBIT¹⁷ (L. Malavolta 2016; L. Malavolta et al. 2016, 2018). The RV signal of each planet was modelled using a Keplerian function. For known transiting planets, the eccentricity was fixed to zero and the argument of pericentre to 90° . Arbitrary tight Gaussian priors were imposed on the orbital period and the time of inferior conjunction around the provided values.

For planets with unknown periods, a range from 10 to 10^2 d was explored in uniform $\log_2$ space, with eccentricity allowed to vary up to 0.5. When the time of inferior conjunction was not available, the mean longitude was fitted instead, assuming a longitude of the ascending node equal to zero. The analysis included up to four additional non-transiting planets alongside the known transiting planet. For planets with known periods, the period and RV semi-amplitude were explored in linear space; otherwise, they were explored in $\log_2$ space.

Stellar activity was mitigated using GP regression through the S+LEAF 2 package (J.-B. Delisle, N. Hara & D. Ségransan 2020; J.-B. Delisle et al. 2022). An Exponential-Sine Periodic kernel (including up to four harmonics) was adopted as a computationally

¹⁴<https://jax.exoplanet.codes/>.

¹⁵<https://num.pyro.ai>.

¹⁶<https://pypi.org/project/numpyro-ext/>.

¹⁷<https://pyorbit.readthedocs.io/>

efficient approximation of the squared-exponential periodic kernel. For the spectroscopic data set, a multivariate GP regression was implemented using FWHM, bisector-span (BIS), equivalent-width (EW), S-index, and $H\alpha$ as activity indicators to constrain the underlying process.

For the photometric data, three scenarios were considered: exclusion of the photometry; GP regression using the same kernel and shared hyperparameters as the spectroscopic data set but with an independent covariance matrix; and full inclusion within the multivariate GP framework, treated analogously to the spectroscopic activity indicators.

The first derivative of the GP was included only for the RV, BIS, and EW time-series (see e.g. G. Mantovan et al. 2024). To avoid degeneracies in the GP coefficients, the convective coefficient associated with the RV was constrained to be positive, while uninformative priors were adopted for all other parameters. Each data set included independent offset and jitter terms. Additional tests considered both unidimensional GP models and configurations using only BIS and S-index as activity indicators.

The parameter space was explored using the affine-invariant MCMC ensemble sampler `emcee` (D. Foreman-Mackey et al. 2013), initialized from the results of the differential evolution optimizer `pyDE` (<https://github.com/hpparvi/PyDE>). The choice of sampler was motivated by the need for controlled execution time. The optimization phase was run for 50 000 generations with `pyDE`, followed by 100 000 MCMC steps with `emcee`, adopting a thinning factor of 100 and a burn-in of 80 000 steps. The number of chains was set dynamically to four times the dimensionality of the model.

Model selection was performed by comparing Bayesian information criterion (BIC) values derived from the RV-only likelihood, with the penalty term computed using only the parameters involved in the RV modelling. This approach was specifically implemented for the RV data challenge.

2.3.5 CCF SCALPELS (A. Collier-Cameron)

This participant applied the SCALPELS algorithm to the cross-correlation functions (CCFs). Since some instrumental systematics can also induce variations in the CCF shape, both the YV0 and YVA realizations of the CCF time-series are analysed.

The autocorrelation function (ACF) of a spectrum (or of a CCF) is invariant under translation. The SCALPELS algorithm (A. Collier Cameron et al. 2021a) exploits this property to identify orthonormal modes of variation in the CCF shape that are insensitive to planetary Doppler shifts. The ACF, denoted $\mathbf{A}$, is computed for each CCF in the time-series using a circular shift and cross-multiplication, with each row normalized by its mean. A singular value decomposition of the resulting time-series is then performed:

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T. \quad (1)$$

The rows of the right-singular matrix $\mathbf{V}^T$ represent orthonormal modes of variation in the ACF shape, while the columns of the left-singular matrix $\mathbf{U}$ provide the corresponding orthonormal time-domain coefficients. The diagonal matrix $\mathbf{S}$ contains eigenvalues ordered by decreasing contribution to the variance of $\mathbf{A}$. In practice, these eigenvalues decrease rapidly, such that most of the signal can be captured by a low-rank reconstruction using only the first few principal components, typically four.

The solar RV time-series contains a small number of outliers, generally arising from anomalous distortions of the CCF that are

poorly described by a low-rank model. These are identified and removed using an iterative scheme based on low-rank reconstruction and outlier rejection.

The shape-driven component of the RV signal is obtained by projecting the RV time-series onto a basis defined by the leading columns of the matrix $\mathbf{U}$. These basis vectors are reordered according to their contribution to the BIC of the RV fit. After this reordering, the first four columns of $\mathbf{U}$ provide a satisfactory reduced-rank representation of the shape-driven RV signal without significant overfitting.

The search for planetary signals is carried out using the KIMA package (<https://github.com/kima-org/kima>; J. P. Faria et al. 2018). This framework simultaneously fits Keplerian orbital models together with linear combinations of the selected columns of $\mathbf{U}$, treated as activity indicators, using the diffusive nested sampling algorithm DNEST4 (B. J. Brewer & C. P. Donovan 2015). Gaussian priors are adopted for the ephemerides of known transiting planets. When only the time of conjunction from a single transit is available, an uninformative prior is used for the orbital period, allowing the algorithm to identify any RV signal consistent with the observed transit time.

The trans-dimensional nature of the sampler allows the number of non-transiting planets to vary up to a maximum value $N_{p,\max}$. Bayesian odds ratios between models with different numbers of planets are derived from their relative occurrence in the posterior samples.

2.3.6 CCF LINSTAR (J. Blanco-Pozo, M. Perger, I. Ribas)

The participants applied the LINSTAR (LINEar Regression for STellar Activity Removal) method, which is an adaptation of the CANSTAR method developed by J. Blanco-Pozo et al. (2026) for this specific challenge. The CANSTAR activity correction relies on training a convolutional attention network on StarSim (E. Herrero et al. 2016) generated data; however, the lack of sufficiently realistic synthetic models of the Sun at the time of the RV challenge led to an adaptation of the method into a purely data-driven approach. The method operated through a two-step procedure:

(i) Separation of line shifts from line-shape distortions in the CCF is achieved through the following decomposition:

$$\begin{aligned} \text{CCF}(\mathbf{v}, t) = & c_0(t) \mathbf{I}(\mathbf{v}) + c_1(t) \mathbf{T}(\mathbf{v}) + l(t) \mathbf{T}'(\mathbf{v}) \\ & + \sum_{n=2}^{N-1} g_n(t) \mathbf{G}_n(\mathbf{v}) + \mathbf{E}(\mathbf{v}, t), \end{aligned} \quad (2)$$

where $\mathbf{I}$ is the unity function, $\mathbf{T}$ is the time-averaged template CCF, $\mathbf{T}' = d\mathbf{T}/d\mathbf{v}$ is its first derivative (capturing the line shift $l(t)$), and $\mathbf{E}$ is the residual error. The functions $\mathbf{G}_n$ form a basis describing line-shape distortions. A Gram–Schmidt orthogonalization is applied to ensure that $\mathbf{I}(\mathbf{v})$, $\mathbf{T}(\mathbf{v})$, and $\mathbf{G}_n$ are strictly orthogonal to $\mathbf{T}'$, enabling a clear separation between Doppler shifts (l) and pure line-shape distortions ($[c_0, c_1, g_2, \dots, g_N]$).

(ii) A joint fit of Keplerian orbits and a linear combination of the selected line-shape distortion coefficients ($[c_0, c_1, g_2, \dots, g_N]$) using the dynamic nested sampler `dynesty` (J. S. Speagle 2020).

For FGK stars, where $\mathbf{T}$ is approximately Gaussian – as in the present study – a Hermite basis is adopted. This basis is constructed from the n -th derivatives of a Gaussian function fitted to $\mathbf{T}$, starting from $n = 2$, since the $n = 0$ and $n = 1$ components correspond to $\mathbf{T}$ and $\mathbf{T}'$. As shown in J. Blanco-Pozo et al. (2026),

these coefficients naturally map onto classical indicators (c_1 to contrast, l to RV, g_2 to FWHM, g_3 to BIS), while extending to higher-order shape variability ($n \geq 4$), which captures both symmetric and asymmetric distortions.

Following an SCALPELS-like strategy (A. Collier Cameron et al. 2021), the optimal number of distortion coefficients is determined iteratively by minimizing the RV root mean square while controlling model complexity through the BIC, thereby avoiding overfitting. Giant planets dominating the time-series variability are pre-whitened via a preliminary least-squares fit prior to this selection.

Once the optimal number of distortion coefficients is established, Keplerian signals are iteratively fitted to the data. For transiting planets, informative priors are applied on periods and phases. For non-transiting planets, uninformative priors are adopted for the orbital phase and semi-amplitude, while period priors are restricted to broad intervals around the dominant peaks in the periodogram to prevent overlap between signals in multi-planet fits.

The dynamic nested sampling is initialized with 500 live points, using a random-walk proposal to explore the parameter space. The sampler computes the log-Bayesian evidence ($\Delta \ln \mathcal{Z}$) by integrating the likelihood over the prior volume, progressively focusing on regions of higher likelihood. The log-likelihood values of the sampled points are stored at each iteration to compute importance weights, which are then used to derive the posterior distributions of the model parameters.

Model complexity is increased iteratively, evaluating both $\Delta \ln \mathcal{Z}$ and the residuals after subtracting the best-fitting model. The addition of further components is halted when no significant signals remain in the residual GLS periodogram (FAP > 1 per cent), or when the increase in model evidence no longer justifies the added complexity ($\Delta \ln \mathcal{Z} \leq 5$; R. Trotta 2008).

2.3.7 SHELL CNN (Y. Zhao)

The SHELL CNN is a convolutional neural network (CNN)-based method designed to model stellar activity signals at the spectral level, thereby improving the detection of Earth-like planets (Y. Zhao et al. 2024). The approach relies on the distinction that planetary signals induce pure Doppler shifts in stellar spectra, while stellar activity produces variations in spectral line profiles, such as changes in asymmetry and depth.

For each data set, the 1D solar spectra are projected into the spectral shell domain ($f_0, \frac{df_0}{d\lambda}$). A solar line mask derived from M. Cretignier et al. (2020a) is used to select strong, unblended spectral lines that are free from telluric contamination.

Next, for each data set, we trained a convolutional neural network from scratch to model the stellar activity-induced RVs, FWHM, and BIS using a 10-fold cross-validation scheme. Specifically, the spectral shell data were randomly divided into 10 groups. The model was then trained over 10 iterations, where in each iteration one group was designated as the test set while the remaining nine groups were used for training. The final fitting results were obtained by concatenating the predictions on the test sets from all iterations.

The trained neural network is used to estimate and subtract the stellar activity-induced RV component from the total RVs, which are measured using a classical CCF applied to the solar spectra. The planetary signal is assumed to follow a circular orbit. A periodogram of the resulting RV residuals is then computed, and a numerically estimated FAP is derived, adopting a

detection threshold of 0.1 per cent. A signal is considered detected if the corresponding peak exceeds this threshold. In such cases, a sinusoidal model is fitted to recover the amplitude and phase of the injected planet. Posterior distributions of the best-fitting Keplerian parameters are subsequently obtained using the U1-traNest package (J. Buchner 2021).

For data sets A1050 and C800, the RV time-series extracted from the solar spectra exhibit strong planetary signals. These dominant signals are first modelled with a single-planet fit and removed prior to training the neural network on the residuals, in order to isolate the stellar activity component.

2.3.8 YARARA V2 (M. Cretignier)

YARARA V2, presented in M. Cretignier et al. (2023), is an extension of YARARA V1 (M. Cretignier et al. 2021). The first version primarily consists of the corrections described in Section 2.2.6, applied directly in the wavelength domain at the spectral level. Once the spectra are corrected, line-by-line (LBL) RVs are extracted following X. Dumusque (2018). Time-domain proxies are then derived by projections onto compact spaces.

The first compact space, called the shell, corresponds to a change in the variables of the spectrum $f(\lambda)$ from the wavelength λ to the flux gradient, $f\left(\frac{df}{d\lambda}\right)$. This transformation, introduced in M. Cretignier et al. (2022), provides two main advantages. First, it preserves information on the normalized flux, which is an important diagnostic of stellar activity as an indirect proxy for photospheric depth (M. Cretignier et al. 2020a; K. Al Moulla et al. 2022; K. Al Moulla et al. 2024). Second, in this representation, a pure Doppler shift has a well-defined form, referred to as the *Doppler shell*, $DS(f, df/d\lambda)$, derived from the first-order approximation of the Doppler shift. A principal component analysis (PCA) is performed in this space to obtain components $PC(f, df/d\lambda)$. The number of components is selected using a modified p -leave-out cross-validation scheme (see appendix of M. Cretignier et al. 2022). A Gram-Schmidt orthogonalization is then applied between these components and the *Doppler shell* before fitting them to the data. The resulting coefficients form time-series that have been shown to correlate with RV variations induced by activity or instrumental systematics.

The second compact space is constructed from the LBL RVs. Under the assumption that a planetary signal corresponds to a pure Doppler shift affecting all spectral lines identically, the variance across LBL RVs is free of planetary contributions (M. Cretignier et al. 2023; M. Ould-Elkhim et al. 2023). Consequently, a PCA performed on LBL RVs, which operates on mean-centred data and captures differential variations between spectral lines, should not contain planetary signals. This again defines a time-domain basis that can be fitted to the data.

YARARA V2 consists of fitting these two time-domain bases to the LBL RVs. Transiting planetary signals are explicitly included in the model to prevent absorption of their amplitudes and to enable the search for additional signals in the residuals. A sinusoidal signal search is performed using an iterative GLS periodogram with a 1 per cent FAP threshold, limited to a maximum of six signals and an orbital period range P between 3 and 5000 d. Once the Keplerian model and the amplitudes of the basis vectors are fixed, the residual RVs are analysed using an MCMC framework, assuming circular orbits and fitting only for the amplitudes of the transiting signals with the DACE platform as for the benchmark in Section 2.3.1.

2.3.9 1D GP (E.C. Gonçalves)

The participant only analysed the data set B1300 due to lack of time. They used the KIMA code (<https://github.com/kima-org/kima>; J. P. Faria et al. 2018) to model the provided RVs with a quasi-periodic kernel. The model included 4 Keplerians for the transiting planets and a quadratic polynomial for the mean trend. The Keplerians had variable period P , K semi-amplitude and T_c time, the prior for the periods was Gaussian with mean being the given period value with a standard deviation of 0.1 (the mono-transiting case was set with a uniform prior from 100 d to 5 times the time span of the time-series), the K semi-amplitude had a uniform prior from 0 to 5 times the variation of the RVs and the prior for the T_c were Gaussian with the mean being the given value and an standard deviation of 0.001 (eccentricities were set to 0 and argument of periastron to $\pi/2$).

A standard quasi-periodic kernel (C. E. Rasmussen & C. K. I. Williams 2006) was used for the GP with the following priors for the hyperparameters η_1 set to default by KIMA (uniform from 0 to the variation of the data max-min), η_2 set with a uniform prior between 15 and 600 d, η_3 set with a uniform prior between 15 and 50, η_4 was left as KIMA's default which is a uniform prior between 0.2 and 5.

The model was run using diffusive nested sampling (B. J. Brewer & D. Foreman-Mackey 2016) for 1 000 000 steps with 16 threads.

3 RESULTS OF THE CHALLENGE

In this section, we present the analysis of the three data sets processed by the participants. We start by defining a score metric (Section 3.1) that can be used in the context of the PLATO mission in order to compare participants (Sections 3.2 and 3.3). Later, the results of all the participants will be put together in order to get broader conclusions for the RV field (Section 3.4).

3.1 Defining a score metric

The output asked to the participants were the posterior samples of the Keplerian fit on the three data sets. We will define a *false detection* as any fitted signal with a period farther than 10 per cent of the injected orbital periods P_{inj} . On the contrary, a *missed signal* will be an injected planetary signal for which no signal was closer than those 10 per cent in period.

We then developed a metric using the semi-amplitude samples K_i to evaluate the methods aiming to closely match the PLATO requirements written in Section 1.1. We note that there is an ambiguity in the requirement; it is unclear whether the 10 per cent figure pertains to the precision or accuracy of the measurement. On the one hand, the accuracy cannot be checked in practice since the ground truth is never known. On the other hand, precise measurements can be biased, such that the confidence in the estimated value is overestimated. The present challenge provides a good benchmark to link the accuracy, which can be computed here since the ground truth is known, to the precision, computable also for real PLATO follow-up data.

In this work, we will define the *accuracy* by $\Delta K/K$ as the relative difference between the injected ground truth K_{inj} and the median of the posterior on the samples K_i :

$$\Delta K/K = \frac{K_{inj} - \text{med}(K_i)}{K_{inj}}. \quad (3)$$

Note that our accuracy definition is signed, but we also use an

unsigned one. It will be clear when signed versus unsigned accuracy was used.

On the other hand, we defined the *precision* σ_K as the sigma width of the posterior, measured by the median absolute deviation (MAD) of the samples, normalized by the ground truth injected value:

$$\sigma_K = \frac{1.48 \times \text{MAD}(K_i)}{K_{inj}}, \quad (4)$$

where the factor 1.48 is just a conversion factor from MAD to a usual standard deviation for a Gaussian distribution.

While it is trivial that neither *accuracy* nor *precision* alone is sufficient to assess the score of a participant, it is certain that a method for which both the *accuracy* and *precision* are good should be considered as successful in general. We therefore defined the *maximum error* ε_K , of the challenge as the worst value between the two:

$$\varepsilon_K = \max(|\Delta K/K|, \sigma_K). \quad (5)$$

Note that with those definitions, the usual ‘number of sigma’ N_σ is just given by the ratio of the *accuracy* over the *precision*:

$$N_\sigma = \frac{|\Delta K/K|}{\sigma_K}. \quad (6)$$

If in practice this *maximum error* ε_K definition is an attempt to obtain a one-dimensional metric to rank the participants, as described later, this ruler is not perfect, and we therefore need to always present the pair $(\Delta K/K, \sigma_K)$ or their alternative representation $(\varepsilon_K, N_\sigma)$ to get a global view of a method.

3.2 Results of the participants

In Fig. 5, we display the K_i posterior distributions for the data set A1050 for all the correctly identified signals for each participant. The question of *false positives* will be discussed later. For later comparison (see Section 3.4), we also created the average posterior from all the participants. Each participant is represented by a unique colour and marker shape displayed on the left and preserved along the paper.

Until the end of the paper, we defined two arbitrary levels as relevant for the PLATO mission. The first level is a 10 per cent level that matches the PLATO requirements, and the second level is a 25 per cent *maximum error* that is the score level needed to achieve accurate planetary atmosphere modelling (S. Seager & D. Deming 2010; J. Crass et al. 2021). We display in Fig. 5, the results satisfying 10 per cent by a green box, and the ones satisfying 25 per cent by a yellow box.

As visible in the figure, for all the signals larger than 50 cm s^{-1} in amplitudes the 10 per cent *maximum error* ε_K is achieved by most participants. However, such a score is not achieved for the HZ transit planets that reach at best the 25 per cent *maximum error*. On a positive note, the transiting planet close to the rotational period of the Sun is recovered at 10 per cent by most participants.

In order to better represent the *accuracy* and *precision*, we extracted the value for each participant and represented all of them in Fig. 6. The same was done for the data sets B1300 and C800, respectively, in Figs 7 and 8. The 10 per cent and 25 per cent levels are now represented by coloured boxes. The score reached by the benchmark is also shown as a black box for comparison.

Such a visualization allows a rapid diagnosis of the easy and challenging planets. For all the data sets, a *maximum error* of 10 per cent was never achieved for any of the HZ planets and only the

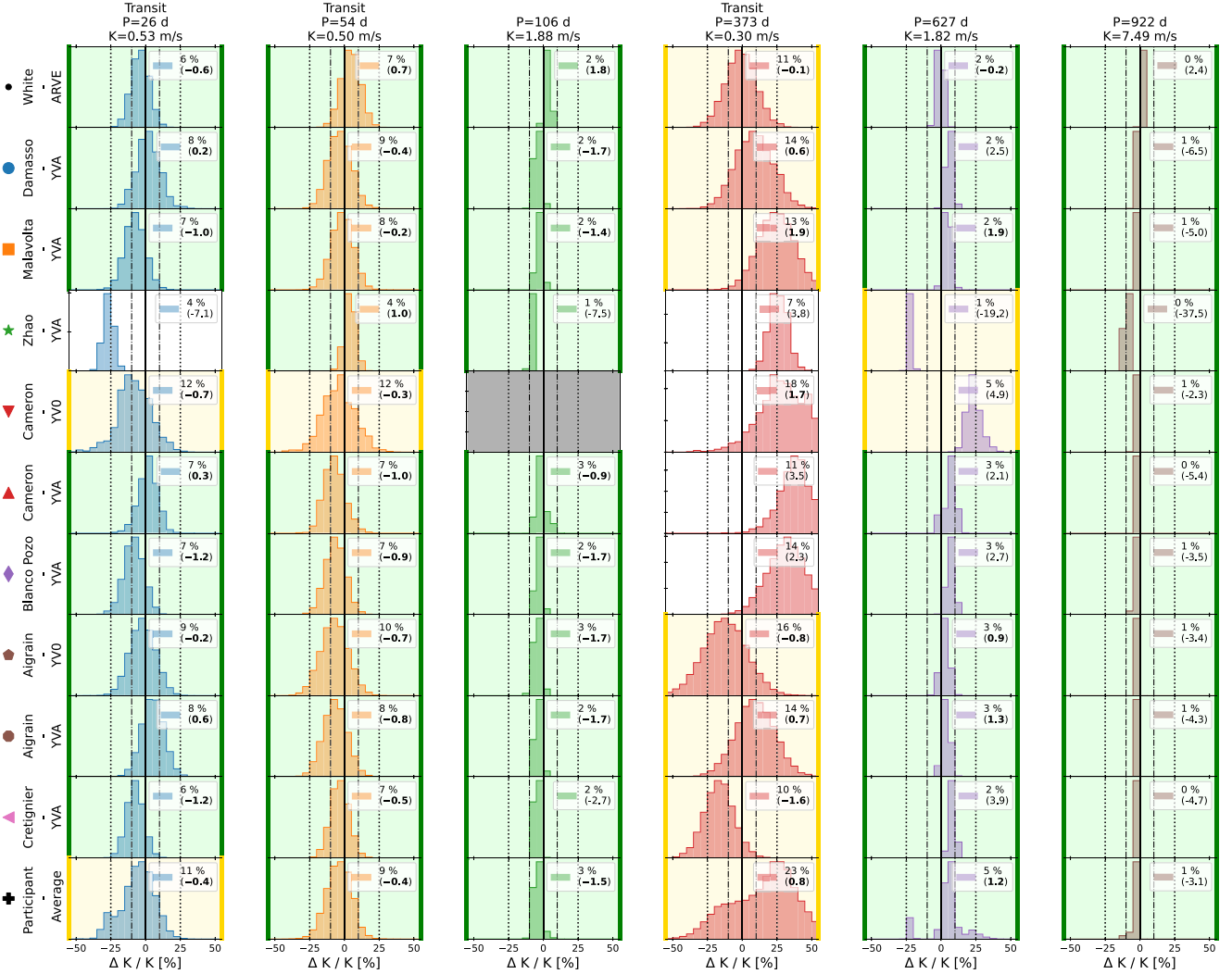


Figure 5. Representation of the K_i semi-amplitude posterior for the six injected planets of the A1050 data set for each participant. Black vertical line is the ground truth K_{inj} , while 10 per cent and 25 per cent levels are indicated by the vertical dash-dotted and dotted lines, respectively. All the panels with a score better than 25 per cent are displayed in a yellow box, while all the scores better than 10 per cent are in green. Planets missed are shown with a black frame. Each panel also indicates the *accuracy* and the number of sigma N_σ . If N_σ is smaller than 2, the value is highlighted in bold. For most participants all the planets are constrained down to 10 per cent. The most difficult signal of the A1050 data set is the HZ transiting signal at 373 d for which no participant gets a 10 per cent score and the mass tends to be overestimated. We do not see any peculiar issue for the transiting planet close to the solar rotation period.

HZ planets of the system A1050 reach the 25 per cent *maximum error*. This confirms the results obtained by N. Meunier, A. M. Lagrange & S. Cuzacq (2019) based on simulations that achieving such a precision level on the mass, even in the case of a transiting planet, is extremely challenging today. Our result even goes one step further, first by using this time real data, secondly by offering a broader range of methods tested. We first discuss below the results obtained on each system individually ordered by number of observations.

3.2.1 Results on C800

This system was certainly the most difficult with semi-amplitudes as small as 20 cm s^{-1} . As visible in Fig. 8, none of the participants reached the 25 per cent level for the transiting planets. On the other hand, most participants correctly recovered the amplitude of the giants, despite the existing cross-talk between the 11-year

gaseous planet and the solar magnetic cycle. The phase of the $2 M_{\text{Jup}}$ signal was slightly anticorrelated with the solar magnetic cycle ($\mathcal{R} = -0.61$) indicating some risk of interaction between the signals that was correctly accounted for by the participants.

The performance on the transiting HZ planet is relatively poor, with a best score of only ~ 30 per cent. However, even the benchmark analysis does not exceed ~ 45 per cent for this signal,¹⁸ indicating an intrinsic difficulty in modelling the data given the available information.

This low performance is likely driven by the proximity of the signal to a one-year period, which results in poor phase coverage.

¹⁸Some participants may occasionally outperform the benchmark, as residual red noise – particularly near a one-year period – can spuriously enhance or suppress the recovered signals depending on its phase relative to the Keplerian signal.

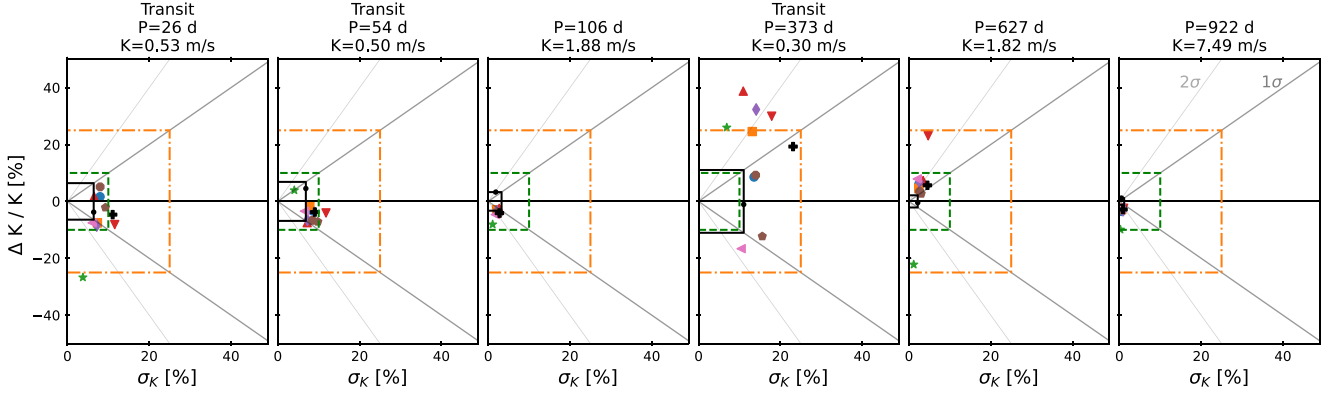


Figure 6. Representation of the pair *accuracy* and *precision* for each injected signal correctly detected by the participants for the data set A1050. Participants are represented with the same symbols as in Fig. 5. The 10 per cent PLATO requirement is indicated with a green dashed square, and the 25 per cent score is shown with the orange dash-dotted square. Any dot inside the square satisfies the score requirement. The black solid box represents the score of the benchmark (see Section 2.3.1). The 1σ and 2σ are shown by the cone plain lines.

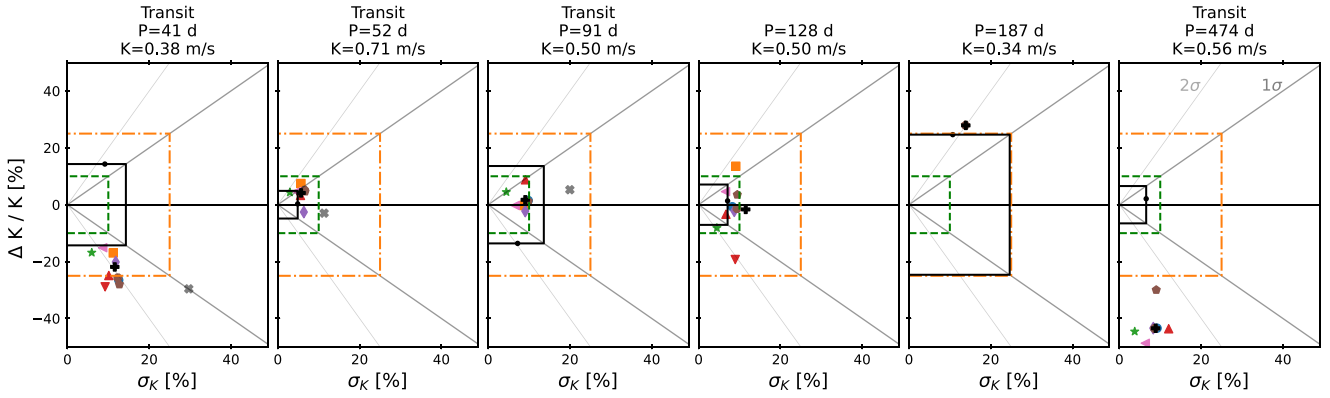


Figure 7. Same as Fig. 6 for the data set B1300. All the participants underestimated the mass of the innermost and outermost planets. If the outermost planet was not constrained since the mono-transit injected case. The underestimation of the inner one is more difficult to understand. No participant except one detected the non-transiting 187-d planet.

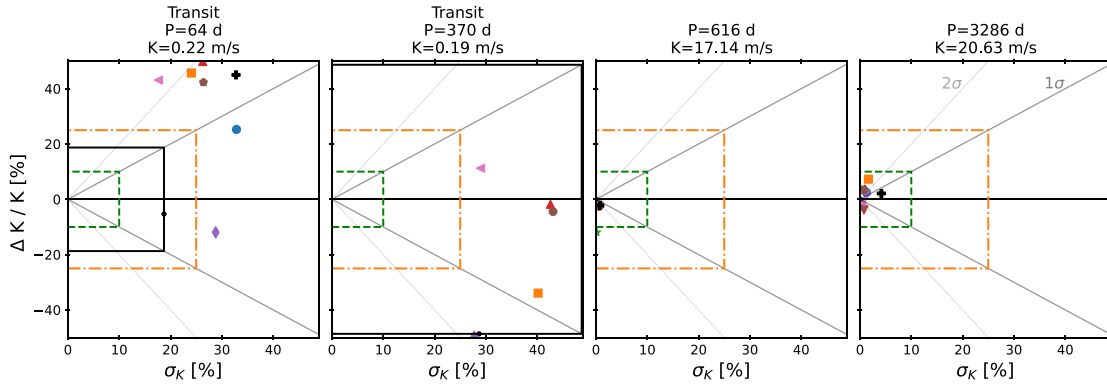


Figure 8. Same as Fig. 6 for the data set C800. Among the smallest amplitudes, no participant manages to reach 25 per cent precision on the transiting terrestrial planets.

This effect is exacerbated by the observational window of C800, which spans only about 7 months, leaving approximately ~ 40 per cent of the orbital phase unobserved.

A more detailed analysis shows that one of the two quadratures is entirely missing. These phases are particularly informative for constraining the semi-amplitude K , and their absence significantly degrades the ability to recover the signal accurately.

The only bad results are obtained for the 64-d transiting signal for which the average participant score is around 40 per cent while the benchmark reaches 20 per cent with a global overestimation of the mass. This likely shows that with realistic data, even 800 measurements (21 nights) are not enough to recover such small amplitude signals even when phases and periods are known.

3.2.2 Results on A1050

The results obtained on this data set are generally speaking the best as displayed in Fig. 6. The level of 25 per cent (and even 10 per cent) is reached several times by the participants. This is notably the case for the 26-d period planet close to the solar rotation period. While it should be reminded that most of our HARPS-N data set is observing a quiet Sun from among the quietest cycles in history (J. Jiang, R. H. Cameron & M. Schüssler 2015; W. D. Pesnell 2016), this seems to confirm that, nonetheless, 50 cm s^{-1} can be relatively well constrained in RVs with an HARPS-like stability precision instrument.

As for C800 (see Section 3.2.1), the worst result is obtained for the 373-d transiting HZ signal with an average score around 25 per cent, while the benchmark grazes the 10 per cent limit. Opposite to the signal in C800, this signal seems better constrained by both the benchmark and the participants. While the presence of 250 new observations (5 extra nights of telescope time) in A1050 compared to C800 should certainly help, it is doubtful that such an improvement can only be obtained by more data and part of the explanation can also be found in the phase coverage of the signal. Indeed, opposite to the signal in C800, the 8-month season length is now placing the seasonal gap no longer on the quadrature of the signal but on the rising section of the phase. The presence of both quadratures certainly helps in the determination of the K semi-amplitude.

Despite the good average score value, this 373-d signal is also the one for which the score of the participants is most fluctuant, spanning between 15 per cent and 40 per cent.

3.2.3 Results on B1300

This data set is the least successful from the challenge. All the participants performed significantly worse than the benchmark, in particular for the mono-transit HZ signal at 474 d, the innermost planet at 41 d, and the 187 d signal almost missed by all the participants. This system definitely contained more complexity than expected.

This result was not anticipated since this data set is the only one for which all the signals are correctly covered in phase, contains the largest number of observations ($N = 1300$ or 33 nights of telescope time) and has the longest season (8 months), so understanding why the signals were not recovered is a key point.

It should be recalled that this HZ planet was the unique mono-transit test performed, meaning that no tight prior on the period was used by the participants, even if this signal was also the most massive transiting HZ of the challenge ($7M_{\oplus}$) and correctly identified by most participants. As an interesting observation, the partial submission participant has attributed the mono-transit to the wrong planets (the 128-d signal), which can be explained by the lower boundary of the period prior spanning down to 100 d.

Furthermore, in this system, almost all the participants used a wrong planetary system model by missing the injected signal at 184 d of 34 cm s^{-1} and included a false detection at 630 d. It is very likely that such a wrong planetary model combined with the broad period prior is at the origin of the poor performance achieved on this data set.

By analysing the system architecture, we can notice a few peculiarities. The first is that three of the signals are close to the one-year harmonics at 183, 122, and 91 d, respectively. While not enough to explain the discrepancy (in particular since no eccentric signals were involved in this challenge), we can imagine

that some one-year systematics interact in some unpredictable way with the planetary signals. Such cross-talk was already mentioned for the previous data set, except that this cross-talk is now occurring on a non-transiting planet.

We can question if the reduction by YARARA producing the YVA data set is not affecting those one-year harmonic signals. This observation is partially motivated by the fact that the only participant that detected the 187-d signal used YV0 and not YVA, but we note that other participants using YV0 also missed this signal though. Furthermore, this explanation would not be coherent with the fact that no significant difference was observed between YV0 and YVA for the systems A and C that were both containing one-year planets.

3.2.4 Global comments

There are several culprits that may explain why no method is today yet able to reach the required precision. For some of the transiting signals, the difficulty may arise due to their proximity with one-year that is an expected very difficult case scenario due to aliasing, poor phase coverage of the signal with seasonal gaps and systematics, all producing strong interaction around the one-year period. Such an issue mainly arises for the HZ of solar-type stars, which can be mitigated when observing cooler K-dwarf for which the HZ moves inwards (see Section 3.5).

Part of the long-period difficulty may be due to imperfect stellar activity correction, even if we tested a rotational period transiting planet scenario. Similarly, supergranulation that is expected to produce part of the jitter observed in the HARPS-N solar data set (K. Al Moulla et al. 2023) may not leak power on such long time-scale except if the properties of the supergranulation may also evolve on long time-scale (N. K. O’Sullivan et al. 2025).

While other exotic stellar activity phenomena may be invoked (such as meridional flows; see e.g. N. Meunier & A. M. Lagrange 2020a), the most likely culprit is the instrument stability itself, as well as granulation and supergranulation. It should be noted that while short-term stability of HARPS-N is known to be better than 50 cm s^{-1} (X. Dumusque et al. 2026), the long-term stability of instruments is very difficult to assess. The result of this challenge may indicate that the long-term stability of HARPS-like instruments is not sufficient to achieve the PLATO requirement for such small signals and more stable spectrographs (such as ESPRESSO) certainly have to be considered or further improvements should be achieved in this domain.

Given the result of the participants, one could infer the level of white noise which would yield the same 1σ uncertainties on the orbital elements (see Table A1). To compute this equivalent white noise, we use the analytical formula giving covariance of a least-squares fit. It has been tested that for the systems at hand, with over 1000 points, this analytical formula is consistent with a full MCMC computation of uncertainties. For systems A and B and the time-series generated with RV, we find that the corresponding level of white noise is 80 cm s^{-1} , in accordance with the simulation, and the YARARA YV2 method yields the same equivalent white noise value, while maintaining a comparable accuracy. In both cases, the amplitude is consistent with the combined effect of p-modes, granulation and super-granulation, the latter one being dominating. To go further, it is therefore extremely important to focus on these signals.

However, a fundamental piece of information to bear in mind is also provided by the benchmark, which also struggles sometimes

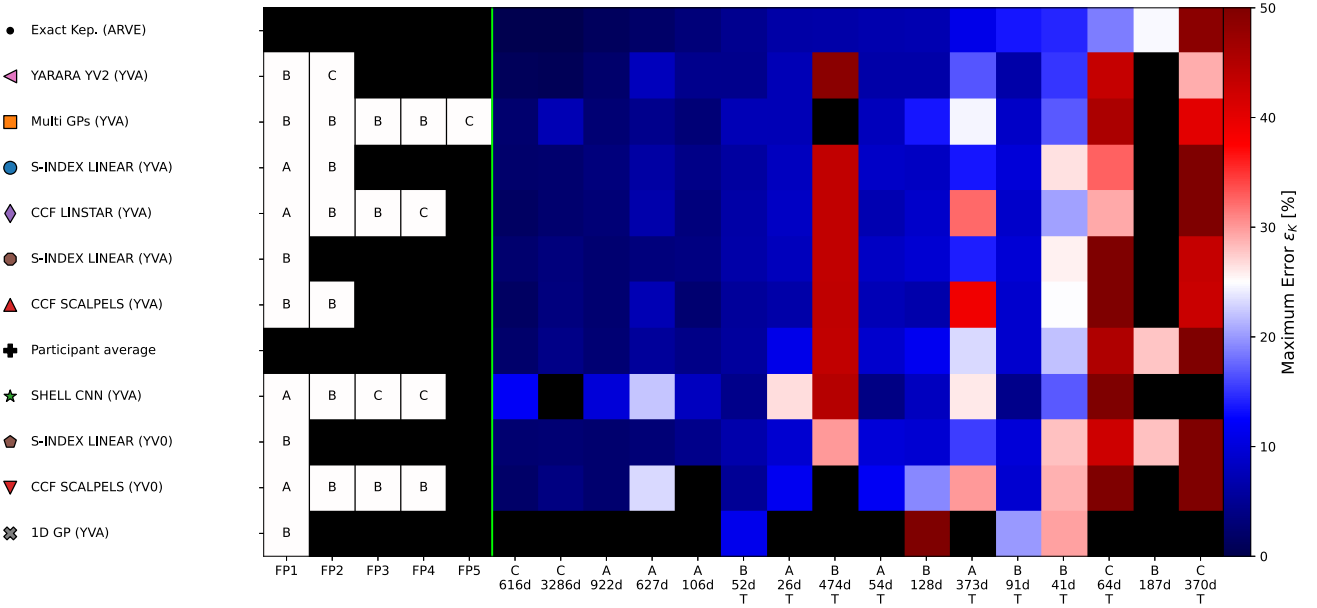


Figure 9. Representation of the *maximum error* ϵ_K obtained by all the participants on each injected planetary signal. Each participant is represented by a row and its unique assigned colour/marker combination. The participants are ordered by their average *maximum error* on the terrestrial signal category, but this ordering does not represent a ranking (see Section 3.3). The exact planetary signal is indicated in the bottom axis label by its system and orbital period (as well as ‘T’ transit indication). The planetary signals were ordered by the *maximum error* obtained by the benchmark (first row). The false positive detections are indicated on the left part of the vertical green line. The scale of the colourbar was set such that the 25 per cent level is white, better scores will be blue, and worse red. A missed signal (true negative) is represented by a black box.

to reach a good score as for the 370-d signal on C800 or 187-d signal in B1300. This seems to indicate that the amount of information in the data was simply insufficient to reach a better level than 25 per cent. Recalling that such a benchmark was obtained using the ARVE solar red noise time-series without rotational modulation or instrumental stability consideration, this indicates that supergranulation alone when combined with suboptimal sampling can forbid such a precision level.

We display in Fig. 9 the *maximum error* ϵ_K obtained by all the participants for each injected planet. The planets (columns) are ordered by the *maximum error* ϵ_K of the benchmark. The participants (rows) are ordered according to their average ϵ_K . By carefully scrutinizing this figure, it becomes clear that the *maximum error* of the participants is in fact closely matching the benchmark even if slightly worse than it where the only exception is the data set B1300.

3.3 Ranking of the methods

There exists no good and unbiased ranking metric to order the participants. First because the ranking may depend on the type of signals, secondly because the ranking metric can be biased by systematic bias in the methodologies used by the participants (see below). We started by categorizing the injected signals into distinct classes. Four classes were defined to evaluate participant performance under different conditions: (1) terrestrial signals, (2) transiting signals, (3) habitable-zone transiting signals, and (4) all the signals.

As an example, we computed for each class the average *maximum error* ϵ_K of all participants¹⁹ and present the results in

Fig. 10. Participants are ranked according to their *maximum error* evaluated over the ‘terrestrial signals’ category (top panel), and this ordering is preserved across the remaining panels for consistency and ease of comparison. While the HZ transiting planets constitute the primary focus of this challenge, they can not be used due to the low statistic number ($N = 3$). The second-best category of reference should have been the transiting planets, but due to the partial submission of one group, this metric is slightly biased for them and we therefore stuck to the terrestrial category for the reference. This choice is without consequence since all the transiting planets are terrestrial and as visible the ordering remains essentially the same using this group or the transiting one.

As expected, the best performance is achieved by the benchmark participant described in Section 2.3.1. With a few exceptions, the ranking of the participants is largely insensitive to the signal category considered. The average *maximum error* ϵ_K is slightly higher when all signals are included, since this subset contains the large gaseous planets (see Fig. 1), which are, by construction, the easiest signals to characterize.

No participant reaches a 10 per cent *maximum error* for the transiting signals, and three methods fall below 25 per cent: (1) YARARA V2, (2) multi-GPs (Malavolta), and (3) linear S-index (Damasso). It is worth noting that even the ARVE benchmark does not reach the 10 per cent level, indicating that such performance was simply not achievable for the injected signals given the number of observations and the adopted realistic noise level.

However, drawing conclusions solely from this global metric would be misleading. In Appendix C, we also present the

¹⁹The grey cross corresponds to a partial submission of the challenge from the participant (Section 2.3.9). Although it cannot be directly compared

to the complete submissions, it is intentionally retained in the figure to highlight the potential risks of drawing conclusions from incomplete results.

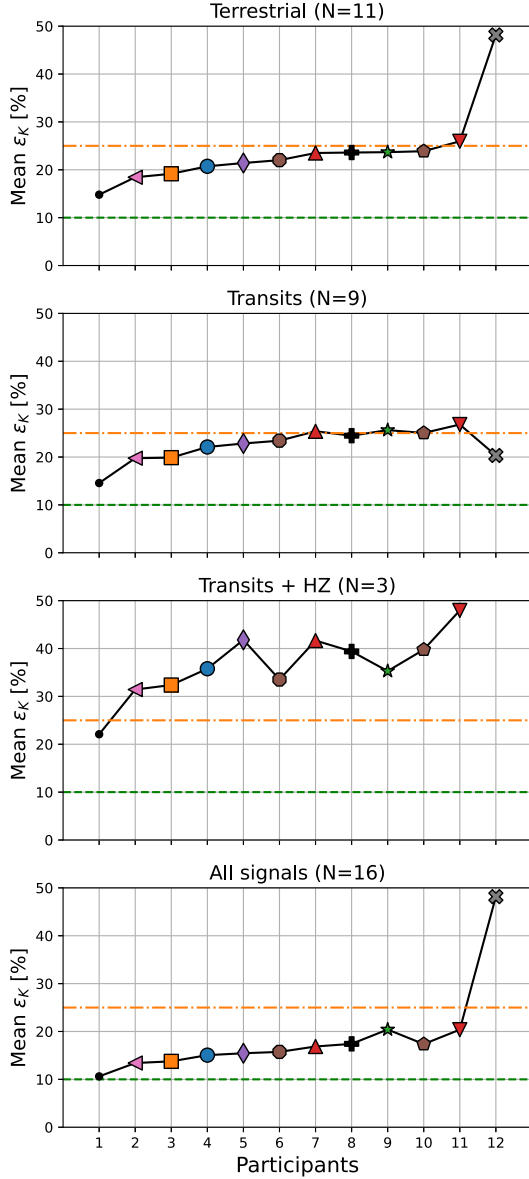


Figure 10. Average *maximum error* ϵ_K for the 16 injected planets. Four different subsets were analysed (from top to bottom panels): terrestrial planets, the transiting planets, the habitable-zone (HZ) transiting planets and the full sample of planets. Each participant is represented by the same marker as in Fig. 9. Participants’ names are however not displayed since this metric alone is not sufficient to allow for a ranking (see the text). The two arbitrary levels defined in this challenge are also displayed as horizontal lines at 10 per cent (green dashed line) and 25 per cent (orange dash-dotted line).

rankings based separately on *precision* σ_K (Fig. C1) and *accuracy* $\Delta K/K$ (Fig. C2). The precision metric is particularly useful for diagnosing methods that underestimate their uncertainties. Such cases can be identified when a method reports a precision better than the benchmark, which effectively represents a near best-case scenario.

Two methods exhibit a better precision than the benchmark analysis: (1) SHELL CNN and (2) YARARA V2. This is suspicious and likely indicates that the uncertainties are underestimated. A plausible explanation is that these methods do not simultaneously fit the Keplerian model and the red-noise component

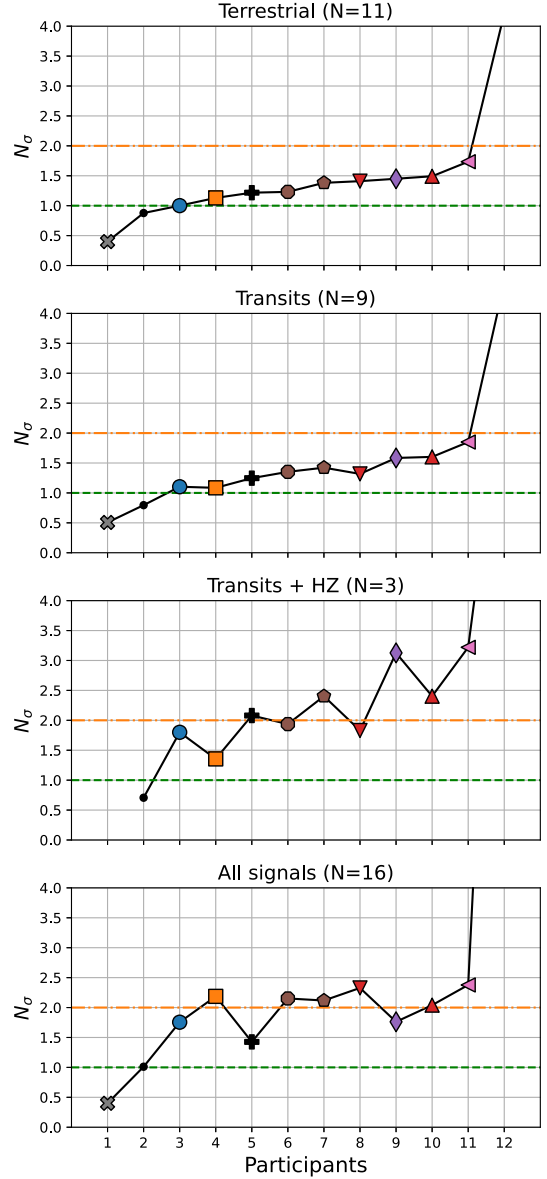


Figure 11. Same as Fig. 10 for the number of sigma N_σ , instead of *maximum error* ϵ_K . The 10 per cent and 25 per cent horizontal levels were replaced by 1σ (dot-dashed line) and 2σ (dotted line) levels that are classical levels. A value above 2σ indicates an underestimation of the uncertainties and overestimation of the precision. A value below 1σ will on the opposite indicate a poor precision.

within a joint MCMC framework. Such rigid models might overlook the degeneracies between model components, which leads to underestimated uncertainties.

Systematic underestimation of uncertainties can bias the *maximum error* ϵ_K ranking. Indeed, a method that overestimates its precision in a systematic way can achieve a better *maximum error* simply because the metric becomes dominated by its *accuracy* while the other participants continue to be evaluated on both metrics. To enable a fair comparison, it is therefore essential to complement the *maximum error* analysis with an evaluation of the number-of-sigma compatibility N_σ for each participant. This analysis is presented in Fig. 11.

As visible by the comparison with Fig. 10, while YARARA V2 was among the best-scoring methods, it is also among the

methods that systematically overestimate their precision with one of the largest N_σ deviation around $N_\sigma = 1.5$ for terrestrial planets and even $N_\sigma \simeq 3.0$ for HZ transiting ones.

In conclusion, we can not rely on a single one-dimensional ruler to decide on the success of a method, and there always subsists a trade-off between accuracy and precision (or their alternative representation). We therefore display the average of each method in the 2D $(\Delta K/K, \sigma_K)$ and $(\varepsilon_K, N_\sigma)$ in Fig. 12. When analysing the methods that have both the smallest maximum error and are the most closely compatible with 1σ two methods can be highlighted: the linear S-index detrending from Section 2.3.3 and the multi-GP from Section 2.3.4. Very interestingly, we see that the ‘best’ participants themselves may depend if the metrics used are the $(\varepsilon_K, N_\sigma)$ from the top row or $(\Delta K/K, \sigma_K)$ from the bottom row.

In definitive terms, we do not clearly attest the existence of an outstanding method to date and all methods seem to compete perfectly on such a data set. This is not extremely surprising given that some method may reveal their plain potential when working on more active stars, or on different S/N data sets.

3.4 Assessing the current RV precision in the RV field

Because this challenge involved a sufficiently large number of participants, employing a wide variety of methodologies – among the most innovative and representative currently available – we attempted to assess the present-day precision of the RV method.

The relatively small number of injected planetary signals is, of course, a limitation and should be kept in mind when interpreting the results, as should the caveats raised earlier regarding the extrapolation of results from the solar data set (see Section 2.2.5).

It is becoming increasingly standard practice not to rely on a single pipeline or analysis when publishing planetary signals, but instead to run several independent pipelines in parallel and compare their outcomes. This healthy practice is now routine for the *James Webb Space Telescope*, where independent pipelines are executed in parallel during data processing (e.g. J. Kirk et al. 2024), and we advocate that a similar standard should progressively become the norm in RV studies.

However, combining results from different reductions or analyses remains a matter of debate. One cannot simply average the outputs of multiple pipelines, since they typically start from (or partially share) the same input data. This makes posterior propagation and model averaging non-trivial.

We therefore adopted a simpler approach: for each injected signal, we selected the best score achieved by any participant. Naturally, this choice introduces a bias and slightly overestimates the final *maximum error* ε_k . As a counterbalance, however, we treated all *false positive* detections as valid detections, even if reported by only a single participant. By doing so, a statistic of 27 false detections is obtained by all the participants combined. The results of this analysis are shown in Fig. 13, and reveal an interesting trend.

Two distinct regions clearly emerge: one where detection scores are high and false positives are essentially absent, and another where scores are low and false detections²⁰ become likely. Although the number of injected planets was insufficient to

densely populate the (K, P) parameter space, a rough empirical boundary $K_{\text{crit}}(P)$ can nevertheless be identified as follows for an HARPS-like instrument:

$$K_{\text{crit}}(P) = \begin{cases} 30 \text{ (cm s}^{-1}\text{)} & P \leq 80 \text{ d} \\ 34 \left(\frac{P}{100 \text{ d}} \right)^{0.634} \text{ (cm s}^{-1}\text{)} & P > 80 \text{ d} \end{cases} \quad (7)$$

The number of false positives emerging from this challenge is a major concern, although they primarily occur – as expected – at low amplitudes and long orbital periods. This result may be concerning since a significant number of planets have been published in this region of the parameter space over the past years. Among the systems falling into the suspicious region are for instance: HD10700 f, HD20794 f, HD134060 c, HD95735 c, HD34445 g, GJ433 c, and GJ15A c.

Although we did not perform a case-by-case reassessment of those published detections, for instance, to verify whether they were supported by independent methods, it is important to emphasize that our challenge was conducted under highly favourable conditions. The data sets typically contain an average of $N \sim 1000$ observations and were based on one of the most stable spectrographs currently available, operating under conditions and S/N that outperform most real stellar RV data sets obtained to date.

Given these considerations, our results suggest that caution is warranted when interpreting low-amplitude, long-period RV detections. In particular, we recommend that planets previously labelled as ‘confirmed’ based solely on RV measurements, especially in this region of parameter space, should again be reconsidered as ‘candidates’ unless supported by independent lines of evidence.

This further strengthens the case for the use of multiple independent pipelines and reduction strategies when assessing RV detections. Cross-validation between methods should become standard practice. However, our results also demonstrate that such sanity checks are not, by themselves, a guarantee against false positives.

In particular, among the false detections identified in the challenge, we find a ~ 600 -d Super-Earth candidate ($\sim 5M_\oplus$) in data set B1300 that was independently detected by all participants. This illustrates that methodological diversity alone does not fully protect against systematic biases or activity-induced signals that can convincingly mimic planetary signatures.

3.5 K-dwarf stars: the RV sweet spot?

In this section, we present a simple scaling analysis of the signals from the RV challenge to assess the impact of spectral type on the PLATO requirements.

Such an exercise should, however, be interpreted with caution. Changing the spectral type affects not only the stellar mass and the location of the HZ, but also the stellar activity budget and the spectral signatures. As a result, it is not straightforward to predict whether the various methods used by the participants would perform better or worse for different spectral types. Moving towards cooler spectral types nevertheless provides several advantages.

First, cooler stars naturally yield larger RV semi-amplitudes for a given planetary mass. In the HZ, the semi-amplitude scales as $K_{\text{HZ}} \propto M_*^{-2/3}$ reflecting both the stellar-mass dependence and the inward shift of the HZ.

Second, the inward shift of the HZ leads to shorter orbital periods, which helps mitigate observational window effects linked

²⁰False detection seems to be ‘overrepresented’ in Fig. 13 that is a graphical bias since all participants produce the same green dots, but unique red dots.

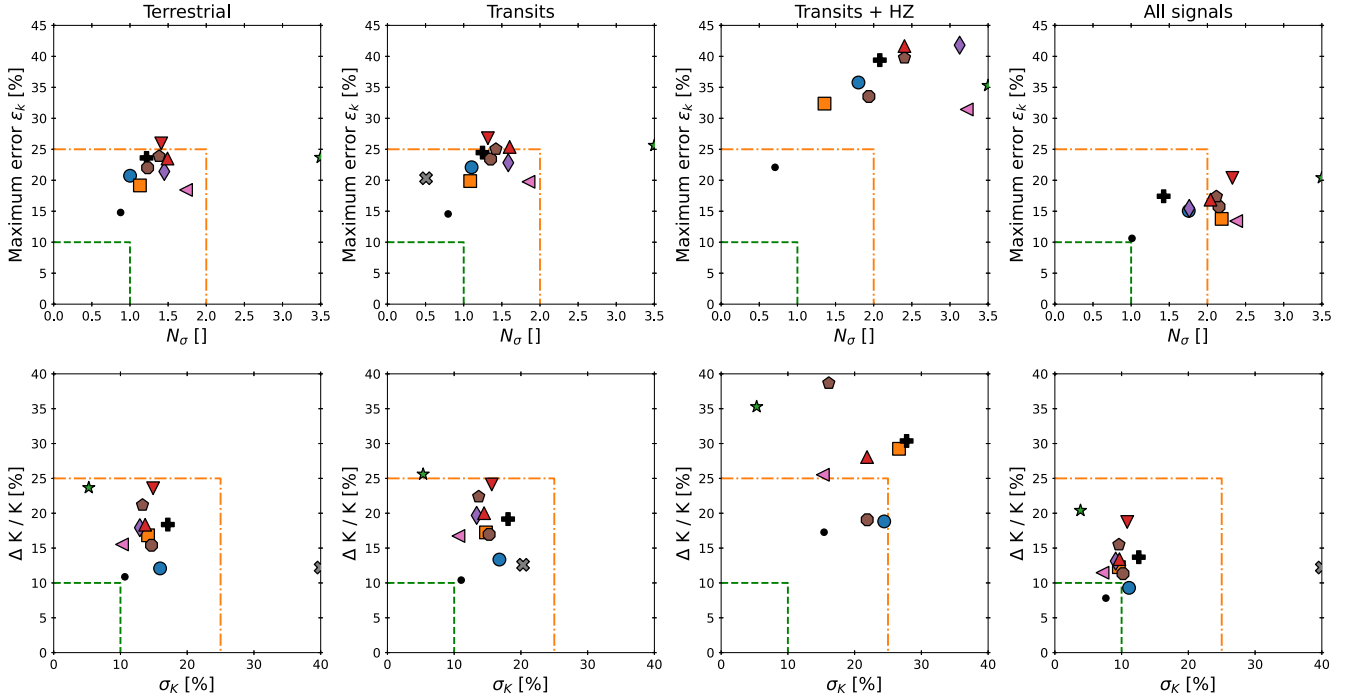


Figure 12. Representation of the four metrics *maximum error* ϵ_k , number of sigma N_σ , absolute *accuracy* $|\Delta K/K|$ and *precision* σ_K obtained for each participant for the four different classes of signals. Top row represents the combination of (N_σ , ϵ_k) and bottom row represents the combination (σ_K , $\Delta K/K$) similarly to Fig. 6.

to Earth’s orbit. This improves phase coverage and therefore enhances both detectability and characterization, as signals can be sampled over multiple orbital cycles.

Finally, the expected RV variability induced by convective flows is reduced. The amplitudes of p-modes, granulation (and supergranulation) are predicted to decrease by about 30 per cent (N. Meunier, A. M. Lagrange & S. Cuzacq 2019a). In addition, cooler stars tend to rotate more slowly and exhibit lower spot contrast. As a consequence, stellar activity signals are expected to be more strongly dominated by the inhibition of convective blueshift (ICB) described in N. Meunier, M. Desort & A.-M. Lagrange (2010), rather than by flux variations induced by active-region contrast (N. Meunier & A.-M. Lagrange 2013). The ICB amplitude is itself smaller since the convective flows are slower for cooler-type stars (F. Liebing et al. 2021; K. Al Moulla et al. 2024).

We illustrate the expected gain in Fig. 14. We adopt the same KP detection limit as the red curve shown in Fig. 12 and defined in equation (7). This limit is then rescaled with stellar mass for two additional spectral types: K1 ($M_* = 0.85, M_\odot$) and K5 ($M_* = 0.70, M_\odot$). The HZ is correspondingly shifted inwards following R. K. Kopparapu et al. (2013, 2014). The resulting HZ period ranges are [272, 800] d for a G2 star, [136, 472] d for a K1 star, and [65, 237] d for a K5 star.

We apply the same scaling to all planetary systems in the challenge by keeping the planetary masses fixed while rescaling their orbital periods such that HZ transiting planets receive the same incident flux as the Earth–Sun system. This analysis reveals that while for a G2-type star the requirements are achieved for $\sim 8 M_\oplus$ planets for the innermost part of the HZ, the PLATO 10 per cent requirements may be fulfilled down to $\sim 3 M_\oplus$ for a K1 and $\sim 2 M_\oplus$ for a K5 star.

We note that similar scaling arguments have previously been used to motivate the search for HZ planets around M dwarfs,

although several important limitations were not anticipated. In particular, M dwarf spectra in the visible are often heavily blended and obtained at a relatively low S/N, which has driven the development of infrared spectrographs. However, the technical challenges associated with infrared observations were initially underestimated (P. Plavchan et al. 2015; D. A. Fischer et al. 2016), and their instrumental stability has only very recently reached the level achieved in the visible (É. Artigau et al. 2024; F. Bouchy et al. 2025; R. Varas et al. 2025).

Furthermore, M dwarfs typically host stronger magnetic fields, leading to enhanced stellar activity signals, in particular through Zeeman broadening (A. Reiners et al. 2013; O. Kochukhov 2021). Finally, the question of habitability itself has been rapidly called into question: the habitable zone lies very close to the host star, and young M dwarfs undergo intense flaring phases that may be sufficient to erode atmospheres or prevent the emergence of life.

4 CONCLUSIONS

In this paper, we presented an RV data challenge organized within the PLATO Stellar Variability Mitigation Working Group to compare different analysis methods and establish a robust analysis strategy in the context of the future PLATO follow-up, in particular regarding the precision and accuracy achievable on RV semi-amplitudes.

The solar HARPS-N data set was used to generate three synthetic data sets designed to closely mimic realistic follow-up observations. Three consecutive 5-min exposures were combined to produce an effective integration time of 15 min, yielding data sets with $S/N_{550} \sim 540$. This S/N corresponds to observations of a star with magnitude $m_v \approx 5.4$ using a 3.6m telescope such as HARPS, and $m_v \approx 7.2$ with ESPRESSO on the VLT.

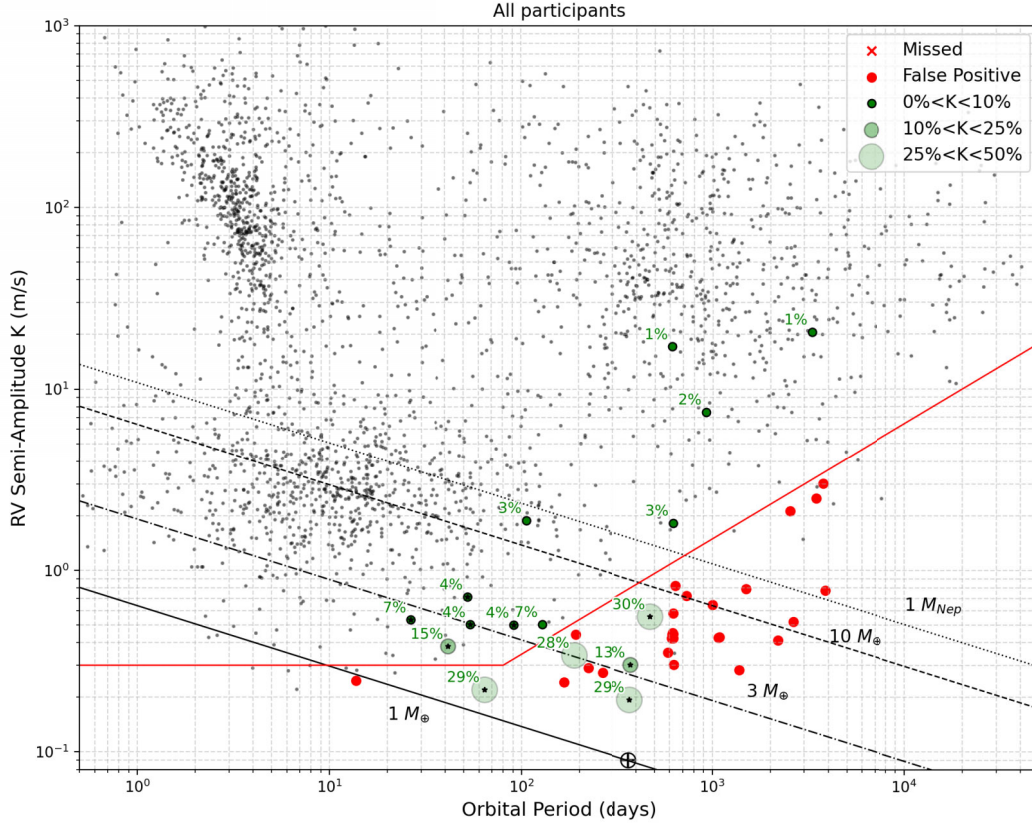


Figure 13. Visualization of the best score obtained by combining the results of all participants using the same (K, P) parameter space as Fig. 1. For each planet, only the best *maximum error* achieved among all participants was retained and is indicated by its value and the marker size (see the legend). As a counterpart, every false positive is also shown, even if detected by a single participant. All injected planetary signals were detected by at least one participant. The red lines delimit the (K, P) parameter-space region where the current RV precision enables reliable planetary detections from the region where published detections should be treated with caution.

Planetary signals were injected directly into the spectra, after which CCFs and RV time-series were derived. Particular care was taken to ensure that the injected signals affected only the stellar component of the spectra, and not systematics tied to the Earth laboratory rest frame.

Three multiplanet systems were simulated, each containing between four and six planets, including one transiting habitable-zone planet as well as additional non-transiting companions. To limit the dimensionality of the problem, all planetary orbits were assumed to be circular, such that only the semi-amplitudes needed to be fitted for the known transiting planets. While this assumption simplifies the analysis, the impact of eccentricity on signal recovery could be explored in future studies.

Four metrics were used to evaluate participant performance based on the recovered K semi-amplitudes: (1) accuracy $\Delta K/K$, (2) precision σ_K , (3) maximum error ε_K , and (4) significance in units of sigma N_σ . No overall ranking was established, as performance was found to depend strongly on both the chosen metric and the specific signals. Moreover, the limited number of injected signals prevents any statistically robust ranking.

For terrestrial-mass signals, no participant achieved a precision better than 10 per cent, and even the benchmark time-series reached only ~ 15 per cent. This indicates that supergranulation alone is sufficient to prevent reaching the PLATO requirement at this level. Most participants exhibit *maximum error* ε_K in the range of 20–25 per cent.

By analysing both false positives and best-performing detections, a region of the (K, P) parameter space was identified where planetary detections should be treated with caution for instruments with HARPS-like stability. Our results highlight the importance of comparing independent analysis methods when claiming low-amplitude planetary signals. In this regime, signals that are not consistently recovered by different methods should currently be considered as planet candidates rather than confirmed RV detections until corroborated by additional evidence.

Additional challenges arise from observational sampling, particularly for signals near a one-year period, which complicates the detection of HZ planets around solar-type (G2) stars. In contrast, for a similar magnitude, K dwarfs benefit from shorter HZ periods. Simple scaling relations suggest that the 10 per cent precision requirement could be met for HZ planets of ~ 8 – 10 , $M_\oplus$ around G2 stars, and down to ~ 2 – 5 , $M_\oplus$ around K dwarfs. These results further strengthen the case for K dwarfs as the most favourable targets for the first robust detections of habitable-zone planets with the RV technique. However, we noticed that this difficulty was enhanced in the RV challenge due to the differential extinction of the Sun producing season length of the three data sets around $S \sim 220$ d, while the visibility of the PLATO south field of view from Paranal is $S \sim 280$ d.

This challenge provides three new benchmark data sets publicly available for the RV community to test and compare analysis methods. It should be noted, however, that future participants

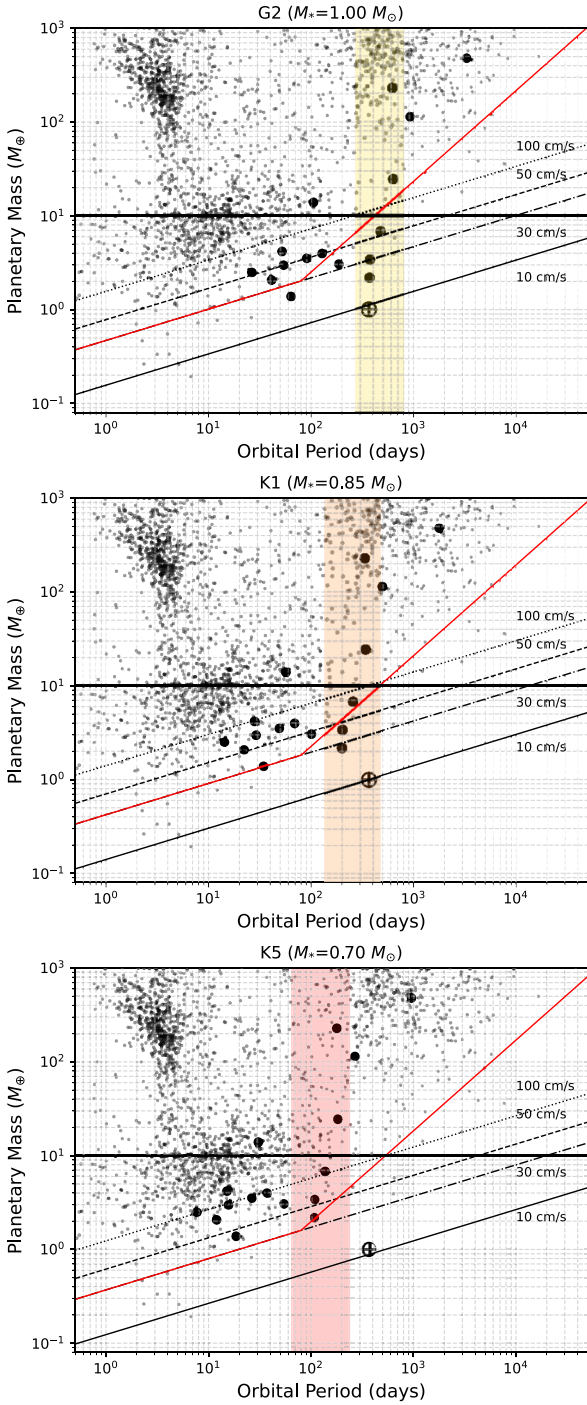


Figure 14. Scalability of the empirical detection limits from the PLATO RV challenge, derived for the Sun (top panel), and extrapolated to a K1 (middle panel) and K5 star (bottom panel) in the *mass-period* diagram. The archival exoplanet data base is displayed for comparison to the signals from this challenge (black markers). The HZ of each star is shown as a shaded region, computed using the conservative limits of R. K. Kopparapu et al. (2013, 2014). The planetary masses from the RV challenge injected planets are kept fixed and periods rescaled to ensure Earth–Sun incident flux. For a G2 solar-type star, only planets with masses $m_p = 10 M_\oplus$ (horizontal thick black line) satisfy the requirements at the inner edge of the HZ. The detection threshold improves for a K1 star, reaching $m_p = 5 M_\oplus$, and further improves significantly for a K5 star, where planets down to $m_p = 2 M_\oplus$ reach confident detection level.

will be aware of the injected systems, whereas this challenge was conducted under near-blind conditions.

Future extensions of this work include investigating the impact of a lower S/N, eccentric signals and incorporating data from additional instruments such as ESPRESSO. The priority seems, to us, to improve the correction of granulation and supergranulation. The latter appears to be one of the dominant limitations for future high-precision RV measurements. A dedicated data challenge focusing specifically on this component would be a natural next step.

Some improvements on the planetary mass determination can be expected for the PLATO mission in particular if transit timing variations are used jointly to the Keplerian fit in RVs (A. Leleu et al. 2024), even if the improvement is yet unclear in the presence of non-transiting planets. The best chance of improvements is expected if the supergranulation is better understood in the coming decade since this component alone is sufficient to make the 10 percent difficult to achieve, in line with the results of N. Meunier & A. M. Lagrange (2020b). Such advancements are expected either thanks to PoET, the solar telescope of ESPRESSO (N. C. Santos et al. 2025) or intensive future RV surveys such as the second Earth project 2ES (J. Stürmer et al. 2024) or the Terra Hunting Experiment (S. J. Thompson et al. 2016; R. D. Hall et al. 2018) both entering operation this year and having supergranulation as part of their open scientific question.

In addition, the long-term stability of high-precision spectrographs should remain a priority, as instrumental interventions and ageing may introduce systematics that become critical on the multiyear time-scales required to detect habitable-zone planets.

ACKNOWLEDGEMENTS

MC and SA acknowledge funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 865624). JBP, IR and MP were funded by Spanish grants PID2021-125627OB-C31, PID2023-150491NB-I00, and PID2024-158486OB-C31, and by ‘ERDF A way of making Europe’, by the programme Unidad de Excelencia María de Maeztu CEX2020-001058-M financed by MCIN/AEI/10.13039/501100011033 and by the MaX-CSIC Excellence Award MaX4-SOMMA-ICE, by the Generalitat de Catalunya/CERCA programme, and by the European Research Council (ERC) under the European Union’s Horizon Europe programme (ERC Advanced Grant SPOTLESS; No. 101140786). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. AM acknowledges funding from a UKRI Future Leader Fellowship, grant number MR/X033244/1 and a UK Science and Technology Facilities Council (STFC) small grant ST/Y002334/1. STH acknowledges support from UKRI-STFC grants ST/X001571/1 and UKRI974. ACC acknowledges support from STFC consolidated grant number ST/V000861/1, UKRI/ERC Synergy Grant EP/Z000181/1 (REVEAL), and UKSA grant number ST/X002217/1. NCS was funded by the European Union (ERC, FIERCE, 101052347). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was supported by Fundação para a Ciência e a

Tecnologia (FCT) through national funds under the research grant UID/04434/2025 (DOI 10.54499/UID/04434/2025). The French researchers acknowledge support on PLATO by the Centre national d'études spatiales (CNES). TGW acknowledges support from the University of Warwick and UKSA. D.N. acknowledges support from PLATO ASI-INAF agreements no. 2022-28-HH.0. KA acknowledges support from the Swiss National Science Foundation (SNSF) under the Postdoc Mobility grant P500PT_230225. NCS, KA, and EG acknowledge funding from the European Union (ERC, FIERCE, 101052347). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through national funds under the research grant UID/04434/2025 (DOI 10.54499/UID/04434/2025). CAW would like to acknowledge support from the UK Science and Technology Facilities Council (STFC, grant number ST/X00094X/1).

This research made use of NASA's Astrophysics Data System (ADS) bibliographic services. We acknowledge the community efforts devoted to the development of the following open-source packages that were used in this work: NUMPY (numpy.org), MATPLOTLIB (matplotlib.org), and ASTROPY (astropy.org). This publication made use of The Data & Analysis Center for Exoplanets (DACE), which is a facility based at the University of Geneva (CH) dedicated to extrasolar planets data visualization, exchange and analysis. DACE is a platform of the Swiss National Centre of Competence in Research (NCCR) PlanetS, federating the Swiss expertise in Exoplanet research. The DACE platform is available at <https://dace.unige.ch>.

DATA AVAILABILITY

The data underlying the RV challenge are publicly available in the Google Drive repository: <https://drive.google.com/drive/u/1/folders/1Jbz-i21D81lR90w3AVSDzygctGuFnqo>. Data are also publicly available at Zenodo ([10.5281/zenodo.105281](https://zenodo.org/record/105281)).

REFERENCES

- Al Moulla K., 2025, *A&A*, 701, A266
 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
 Al Moulla K., Dumusque X., Cretignier M., 2024, *A&A*, 683, A106
 Anna John A. et al., 2025, *MNRAS*, 543, 1974
 Artigau É. et al., 2024, in Bryant J. J., Motohara K., Vernet J. R. D., eds, *Proc. SPIE Conf. Ser. Vol. 13096, Ground-based and Airborne Instrumentation for Astronomy X*. SPIE, Bellingham, p. 130960C
 Baluev R. V., 2015, *MNRAS*, 446, 1478
 Batalha N. M., 2014, *Proc. Natl. Acad. Sci.*, 111, 12647
 Blanco-Pozo J. et al., 2026, *A&A*, 710, A194
 Borucki W. J. et al., 2010, *Science*, 327, 977
 Bouchy F. et al., 2025, *A&A*, 700, A10
 Brewer B. J., Donovan C. P., 2015, *MNRAS*, 448, 3206
 Brewer B. J., Foreman-Mackey D., 2016, preprint ([arXiv:1606.03757](https://arxiv.org/abs/1606.03757))
 Bryson S., Coughlin J., Batalha N. M., Berger T., Huber D., Burke C., Dotson J., Mullally S. E., 2020, *AJ*, 159, 279
 Buchner J., 2021, *J. Open Source Softw.*, 6, 3001
 Buchner J. et al., 2014, *A&A*, 564, A125
 Burke C. J. et al., 2015, *ApJ*, 809, 8
 Cabrera J. et al., 2026, preprint ([arXiv:2604.04818](https://arxiv.org/abs/2604.04818))
 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., 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., 2024a, *MNRAS*, 527, 2940
 Cretignier M. et al., 2024b, *MNRAS*, 535, 2562
 Cui K., Armstrong D. J., Hadjigeorgiou A., Lafarga M., Kunovac V., Doyle L., Nieto L. A., Díaz R. F., 2026, *MNRAS*, 546, stag022
 de Beurs Z. L. et al., 2022, *AJ*, 164, 49
 Delisle J.-B., Hara N., Ségransan D., 2020, *A&A*, 638, A95
 Delisle J.-B., Unger N., Hara N. C., Ségransan D., 2022, *A&A*, 659, A182
 Dravins D., Ludwig H.-G., 2024, *A&A*, 687, A60
 Dumusque X., 2018, *A&A*, 620, A47
 Dumusque X. et al., 2015, *ApJ*, 814, L21
 Dumusque X. et al., 2017, *A&A*, 598, A133
 Dumusque X. et al., 2021, *A&A*, 648, A103
 Dumusque X. et al., 2026, *A&A*, 706, A231
 Faria J. P., Santos N. C., Figueira P., Brewer B. J., 2018, *J. Open Source Softw.*, 3, 487
 Feroz F., Balan S. T., Hobson M. P., 2011, *MNRAS*, 415, 3462
 Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, *Open J. Astrophys.*, 2, 10
 Figueira P. et al., 2025, *A&A*, 700, A174
 Fischer D. A. et al., 2016, *PASP*, 128, 066001
 Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
 Fulton B. J., Petigura E. A., Blunt S., Sinukoff E., 2018, *PASP*, 130, 044504
 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
 Grouffal S. et al., 2025, *A&A*, 701, A173
 Hall R. D., Thompson S. J., Handley W., Queloz D., 2018, *MNRAS*, 479, 2968
 Hara N. C., Ford E. B., 2023, *Annu. Rev. Stat. Appl.*, 10, 623
 Hara N. C., Boué G., Laskar J., Delisle J.-B., Unger N., 2019, *MNRAS*, 489, 738
 He M. Y., Ford E. B., Ragozzine D., 2021, in *Posters from the TESS Science Conference II (TSC2)*. Zenodo, p. 122
 Herrero E., Ribas I., Jordi C., Morales J. C., Perger M., Rosich A., 2016, *A&A*, 586, A131
 Jannsen N. et al., 2024, *A&A*, 681, A18
 Jiang J., Cameron R. H., Schüssler M., 2015, *ApJ*, 808, L28
 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
 Kirk J. et al., 2024, *AJ*, 167, 90
 Klein B. et al., 2024, *MNRAS*, 531, 4238
 Klein B. et al., 2025, *MNRAS*, 542, 2714
 Koch D. G. et al., 2010, *ApJ*, 713, L79
 Kochukhov O., 2021, *A&AR*, 29, 1
 Kopparapu R. K. et al., 2013, *ApJ*, 765, 131
 Kopparapu R. K., Ramirez R. M., SchottelKotte J., Kasting J. F., Domagal-Goldman S., Eymet V., 2014, *ApJ*, 787, L29
 Leleu A. et al., 2024, *A&A*, 688, A211

- Liebing F., Jeffers S. V., Reiners A., Zechmeister M., 2021, *A&A*, 654, A168
- Lienhard F., Mortier A., Buchhave L., Collier Cameron A., López-Morales M., Sozzetti A., Watson C. A., Cosentino R., 2022, *MNRAS*, 513, 5328
- Lissauer J. J. et al., 2011, *Nature*, 470, 53
- Luque R. et al., 2023, *Nature*, 623, 932
- Malavolta L., 2016, Astrophysics Source Code Library, record ascl:1612.008
- Malavolta L. et al., 2016, *A&A*, 588, A118
- Malavolta L. et al., 2018, *AJ*, 155, 107
- Mantovan G. et al., 2024, *A&A*, 682, A129
- Mayor M. et al., 2003, *The Messenger*, 114, 20
- Meunier N., Lagrange A.-M., 2013, *A&A*, 551, A101
- Meunier N., Lagrange A.-M., 2019, *A&A*, 628, A125
- Meunier N., Lagrange A. M., 2020a, *A&A*, 638, A54
- Meunier N., Lagrange A. M., 2020b, *A&A*, 642, A157
- 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., Lagrange A. M., Cuzacq S., 2019, *A&A*, 632, A81
- Montalto M. et al., 2026, preprint (arXiv:2604.03369)
- Nascimbeni V. et al., 2026, *A&A*, 710, A300
- O’Sullivan N. K. et al., 2025, *MNRAS*, 541, 3942
- Oranje B. J., 1983, *A&A*, 124, 43
- Ould-Elhkim M. et al., 2023, *A&A*, 675, A187
- Pepe F., Ehrenreich D., Meyer M. R., 2014, *Nature*, 513, 358
- Pepe F. et al., 2021, *A&A*, 645, A96
- Pesnell W. D., 2016, *Space Weather*, 14, 10
- Petigura E. A., Howard A. W., Marcy G. W., 2013, *Proc. Natl. Acad. Sci.*, 110, 19273
- Plavchan P. et al., 2015, preprint (arXiv:1503.01770)
- Plotnikov M., Valencia D., 2024, *MNRAS*, 530, 3488
- Rasmussen C. E., Williams C. K. I., 2006, *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA
- Rauer H. et al., 2014, *Exp. Astron.*, 38, 249
- Rauer H., Heras A., Mas-Hesse M., Pagano I., 2024, in *European Planetary Science Congress*. p. EPSC2024–197
- Rauer H. et al., 2025, *Exp. Astron.*, 59, 26
- Reiners A., Shulyak D., Anglada-Escudé G., Jeffers S. V., Morin J., Zechmeister M., Kochukhov O., Piskunov N., 2013, *A&A*, 552, A103
- Ricker G. R. et al., 2015, *J. Astron. Telesc. Instrum. Syst.*, 1, 014003
- Royer R. L., Steffen J. H., 2025, *Open J. Astrophys.*, 8, 145
- Santerne A. et al., 2019, preprint (arXiv:1911.07355)
- Santos N. C. et al., 2025, preprint (arXiv:2505.08540)
- 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
- Seager S., Deming D., 2010, *ARA&A*, 48, 631
- Seifahrt A., Stürmer J., Bean J. L., Schwab C., 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. 107026D
- Shahaf S., Zackay B., 2023, *MNRAS*, 525, 6223
- Shen Y., Turner E. L., 2008, *ApJ*, 685, 553
- Silva A. M. et al., 2022, *A&A*, 663, A143
- Silva A. M. et al., 2025, *A&A*, 700, A93
- Speagle J. S., 2020, *MNRAS*, 493, 3132
- Stalport M. et al., 2023, *A&A*, 678, A90
- Stürmer J. et al., 2024, in Bryant J. J., Motohara K., Vernet J. R. D., eds, *Proc. SPIE Conf. Ser. Vol. 13096, Ground-based and Airborne Instrumentation for Astronomy X*. SPIE, Bellingham, p. 130968E
- Thompson A. P. G. et al., 2020, *MNRAS*, 494, 4279
- 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, p. 99086F
- Trotta R., 2008, *Contemp. Phys.*, 49, 71
- Vanderburg A. et al., 2016, *ApJ*, 827, L10
- Varas R. et al., 2025, *Exp. Astron.*, 60, 10
- Wells R., Poppenhaeger K., Watson C. A., Heller R., 2018, *MNRAS*, 473, 345
- Yu H., Aigrain S., Klein B., Cretignier M., Lienhard F., Roberts S. J., 2024, *MNRAS*, 535, 634
- Zechmeister M., Kürster M., 2009, *A&A*, 496, 577
- Zechmeister M. et al., 2018, *A&A*, 609, A12
- Zeng L., Sasselov D. D., Jacobsen S. B., 2016, *ApJ*, 819, 127
- Zhao J., Tinney C. G., 2020, *MNRAS*, 491, 4131
- Zhao J., Ford E. B., Tinney C. G., 2022, *ApJ*, 935, 75
- Zhao L. L., Hogg D. W., Bedell M., Fischer D. A., 2021, *AJ*, 161, 80
- Zhao L. L. et al., 2022, *AJ*, 163, 171
- Zhao L. L., Bedell M., Hogg D. W., Luger R., 2024, *ApJ*, 977, 140
- Zhao Y. et al., 2024, *A&A*, 687, A281

APPENDIX A: TABLES OF PARTICIPANTS OUTPUT

Table A1. K_1 semi-amplitudes posterior distributions of the participants for the planets in the challenge as well as false positive detections. The values represent the median of the distribution and the uncertainties estimated with 1.48 times the MAD. We indicated the false positive (FP) with their period in days in parentheses. All the values for the K semi-amplitudes are in cm s^{-1} .

Planets	Injected	Exact Kep. (ARVE)	S-index linear (YVA)	Multi- GPs (YVA)	SHELL CNN (YVA)	CCF SCALPELS (YVA)	CCF SCALPELS (YV0)	CCF LINSTAR (YVA)	S-index linear (YVA)	S-index linear (YV0)	YARARA YV2 (YVA)	1D GP (YVA)	Participant average
A26d	53	55 $\pm$ 3	54 $\pm$ 4	57 $\pm$ 4	68 $\pm$ 2	54 $\pm$ 3	58 $\pm$ 6	58 $\pm$ 4	56 $\pm$ 4	54 $\pm$ 5	57 $\pm$ 3		56 $\pm$ 6
A54d	50	53 $\pm$ 3	52 $\pm$ 4	51 $\pm$ 4	52 $\pm$ 2	54 $\pm$ 4	52 $\pm$ 6	54 $\pm$ 4	54 $\pm$ 4	54 $\pm$ 5	52 $\pm$ 3		52 $\pm$ 5
A106d	188	195 $\pm$ 3	196 $\pm$ 5	194 $\pm$ 4	204 $\pm$ 2	193 $\pm$ 5		195 $\pm$ 4	196 $\pm$ 4	197 $\pm$ 5	197 $\pm$ 3		196 $\pm$ 5
A373d	30	30 $\pm$ 3	33 $\pm$ 4	38 $\pm$ 4	38 $\pm$ 2	42 $\pm$ 3	39 $\pm$ 5	40 $\pm$ 4	33 $\pm$ 4	34 $\pm$ 5	35 $\pm$ 3		36 $\pm$ 7
A627d	182	183 $\pm$ 4	193 $\pm$ 4	191 $\pm$ 4	223 $\pm$ 2	196 $\pm$ 6	224 $\pm$ 9	195 $\pm$ 5	189 $\pm$ 5	187 $\pm$ 5	197 $\pm$ 4		193 $\pm$ 8
A922d	749	757 $\pm$ 4	775 $\pm$ 4	770 $\pm$ 4	823 $\pm$ 2	768 $\pm$ 4	767 $\pm$ 8	772 $\pm$ 7	769 $\pm$ 5	767 $\pm$ 5	765 $\pm$ 4		770 $\pm$ 7
FP1			41 (2185d)		29 (222d)		250 (3465d)	52 (2626d)					
B41d	38	44 $\pm$ 3	48 $\pm$ 5	45 $\pm$ 4	45 $\pm$ 2	48 $\pm$ 4	49 $\pm$ 4	46 $\pm$ 5	48 $\pm$ 5	49 $\pm$ 5	44 $\pm$ 3	49 $\pm$ 11	46 $\pm$ 4
B52d	71	72 $\pm$ 3	75 $\pm$ 4	77 $\pm$ 4	74 $\pm$ 2	74 $\pm$ 4	74 $\pm$ 4	73 $\pm$ 5	75 $\pm$ 5	75 $\pm$ 5	75 $\pm$ 3	74 $\pm$ 8	74 $\pm$ 4
B91d	50	57 $\pm$ 4	51 $\pm$ 5	50 $\pm$ 4	52 $\pm$ 2	54 $\pm$ 5	50 $\pm$ 5	51 $\pm$ 5	50 $\pm$ 5	51 $\pm$ 5	50 $\pm$ 3	53 $\pm$ 10	51 $\pm$ 5
B128d	50	51 $\pm$ 4	51 $\pm$ 4	57 $\pm$ 5	54 $\pm$ 2	52 $\pm$ 3	60 $\pm$ 5	51 $\pm$ 4	51 $\pm$ 5	52 $\pm$ 5	53 $\pm$ 3	56 $\pm$ 66	51 $\pm$ 6
B187d	34	42 $\pm$ 4								44 $\pm$ 5			44 $\pm$ 5
B474d	56	57 $\pm$ 4	80 $\pm$ 5		80 $\pm$ 2	80 $\pm$ 7		80 $\pm$ 5	80 $\pm$ 5	72 $\pm$ 5	83 $\pm$ 4		80 $\pm$ 5
FP1			44 (620d)	25 (14d)	42 (627d)	42 (612d)	82 (642d)	45 (621d)	45 (624d)	58 (626d)	30 (630d)	43 (1073)	
FP2				35 (586d)		28 (1381d)	212 (2528d)	24 (166d)					
FP3				79 (1498d)			301 (3768d)						
FP4				77 (3860d)									
C64d	22	23 $\pm$ 4	28 $\pm$ 7	32 $\pm$ 5	39 $\pm$ 2	33 $\pm$ 6	33 $\pm$ 6	25 $\pm$ 6	33 $\pm$ 6	31 $\pm$ 6	32 $\pm$ 4		32 $\pm$ 7
C370d	19	29 $\pm$ 6	20 $\pm$ 10	26 $\pm$ 8		20 $\pm$ 8	32 $\pm$ 6	29 $\pm$ 5	20 $\pm$ 8	34 $\pm$ 5	22 $\pm$ 6		25 $\pm$ 10
C616d	1714	1718 $\pm$ 4	1754 $\pm$ 11	1755 $\pm$ 5	1919 $\pm$ 2	1742 $\pm$ 8	1750 $\pm$ 7	1744 $\pm$ 14	1755 $\pm$ 6	1759 $\pm$ 5	1733 $\pm$ 5		1752 $\pm$ 13
C3286d	2063	2070 $\pm$ 6	2114 $\pm$ 26	2213 $\pm$ 34		2139 $\pm$ 20	2142 $\pm$ 16	2115 $\pm$ 27	2134 $\pm$ 13	2120 $\pm$ 12	2081 $\pm$ 6		2107 $\pm$ 86
FP1				27 (264d)	44 (190d)			43 (1090d)			64 (1000d)		
FP2					72 (733d)								

APPENDIX B: EVALUATION OF THE SPECTRA RV SHIFT INJECTION ON THE CCF MOMENTS

A major update of this RV challenge compared to previous efforts is the injection of planetary signals directly into the spectra. To implement this, we first evaluated the impact of the RV-shift injection algorithm in order to validate its use for the challenge.

In principle, the cleanest way to inject a planetary signal is to apply a Doppler shift to the wavelength solution prior to computing the CCFs using the standard extraction algorithm. However, this approach was not compatible with our framework, as the RV challenge relies on YARARA to process and clean the spectral time-series.

YARARA operates on a spectral time-series matrix in which all spectra share a common wavelength grid. This uniform

grid is a core assumption of the data format required by the code. Consequently, any injected RV shift must be followed by a reinterpolation step back onto the native wavelength grid.

While reinterpolation is sometimes regarded with caution in high-resolution spectroscopy, there is no strong evidence that it significantly degrades the data quality when applied carefully. In the worst case, interpolation may introduce a slight increase in the noise level, but this effect is typically negligible and does not materially affect the scientific interpretation.

In our implementation, the shifted spectra were reinterpolated onto the original grid using cubic splines. Naturally, the validity of such an interpolation depends on the amplitude of the Doppler shift. What constitutes a ‘small’ shift must therefore be evaluated in the specific context of the challenge.

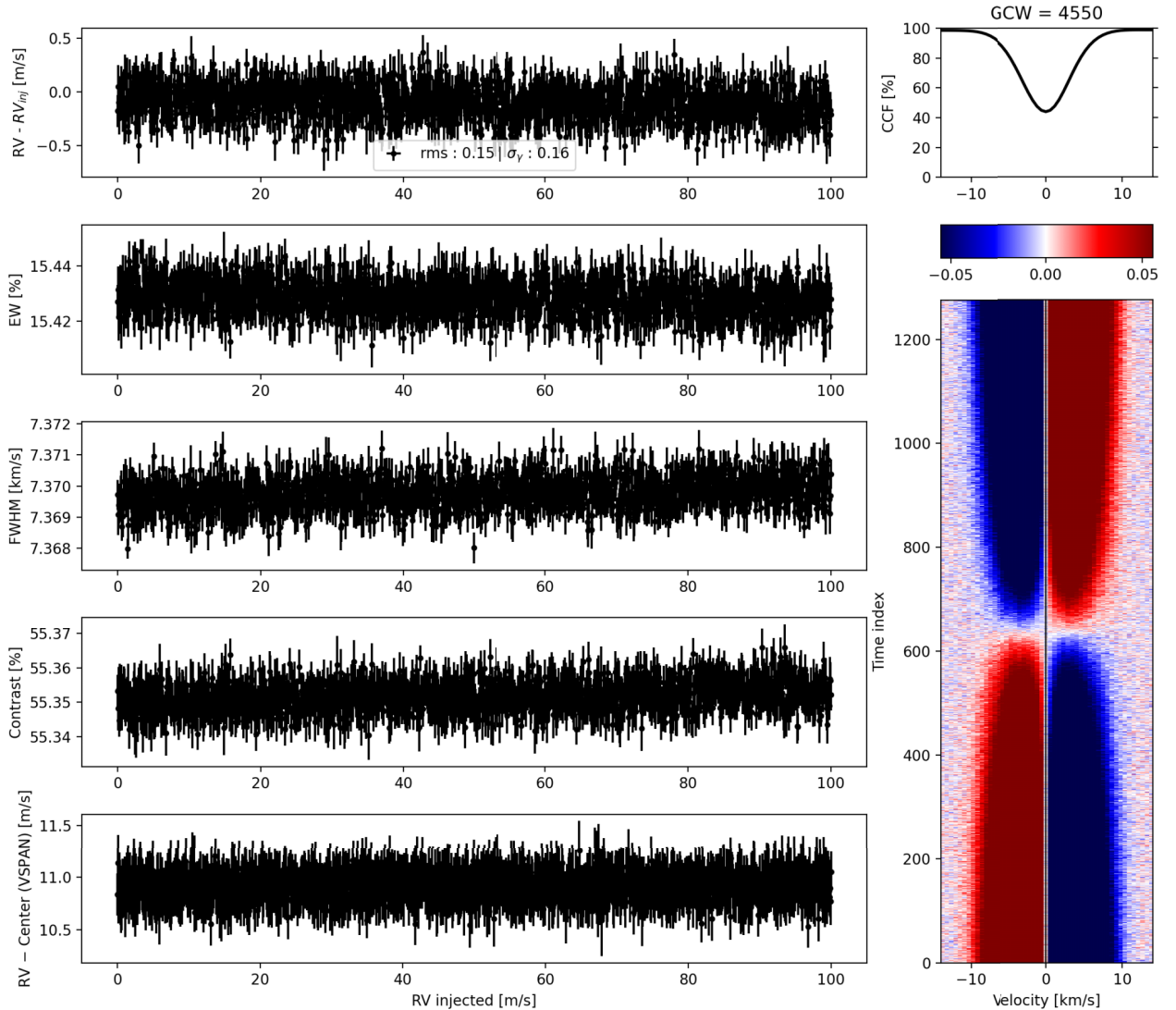


Figure B1. Evaluation of the RV shift injection method used in this RV challenge. The same spectrum was duplicated and shifted on a grid of RV shifts between 0 and 100 m s^{-1} . Photon noise injection was done independently on each spectrum to mimic the $\sigma_{\text{RV}} \simeq 15 \text{ cm s}^{-1}$ RV precision of the challenge. The CCFs are displayed in the top right corner. The residual of the CCFs with the time median CCF is shown in the right heat map, while the different CCF moments, as fitted from a Gaussian fit, are displayed on the different horizontal panels. No significant deviation is observed between the injected RVs and the recovered ones for all the tested velocity shifts. Uncertainties are the propagated photonic flux uncertainties and match the intrinsic dispersion. A small leak of the RV shift interpolation is visible in the FWHM (third panel) as visible by the small linear trend. However, the RV shift was restricted to 20 m s^{-1} .

To quantify the impact of the method, we selected a representative spectrum from the data set and created an exact copy. Photon noise was then injected to reproduce the final RV precision of the solar data set used in the challenge, namely $\sigma_{\text{RV}} \simeq 15 \text{ cm s}^{-1}$. We applied Doppler shifts ranging from 0 to 100 m s^{-1} , followed by reinterpolation, exactly as performed in the construction of the RV challenge.

The results are shown in Fig. B1. We did not observe any significant deviation between the injected and recovered RV values over the full tested range up to 100 m s^{-1} . However, a weak linear trend is visible in the full width at half-maximum (FWHM) of the CCF, indicating a small leakage of the Doppler shift into the line-profile shape.

We quantified this leakage by fitting a linear trend to the CCF moments. For the RV itself, we obtained:

$$\text{RV} - \text{RV}_{\text{inj}} = -0.172 \times \text{RV}_{\text{inj}} \left(\frac{\text{cm s}^{-1}}{\text{m s}^{-1}} \right).$$

This implies that for an injected shift of 100 m s^{-1} , the recovered RV is underestimated by 17 cm s^{-1} . This is a relative difference of 0.172 per cent on the injected amplitude, far smaller than the 10 per cent objective of this challenge. Such a bias is extremely small, which demonstrates that the method selected here will not introduce any significant bias for the participants and confirms the validity of the method. However, given the precision requirements of the challenge, we therefore limited the Keplerian semi-amplitudes to 20 m s^{-1} , ensuring that the maximum bias remains below 3.4 cm s^{-1} .

Furthermore, restricting the injected amplitudes to this range also prevents significant contamination of the CCF moments, since over such RV ranges the leakage remains below the intrinsic jitter precision of the data set for all considered CCF diagnostics.

APPENDIX C: RANKING OF THE PARTICIPANTS WITH PRECISION AND ACCURACY

While an attempt was made to develop a single one-dimensional ruler to rank the participants, for instance by the use of the *maximum error* ϵ_k metric that combines both the *accuracy* and *precision*, we realise that all the score metrics are imperfect to some degree and still require a 2D parameter space to analyse the results.

We display the ranking of all the participants based on each individual metric, respectively, the *precision* in Fig. C1 and the *accuracy* in Fig. C2.

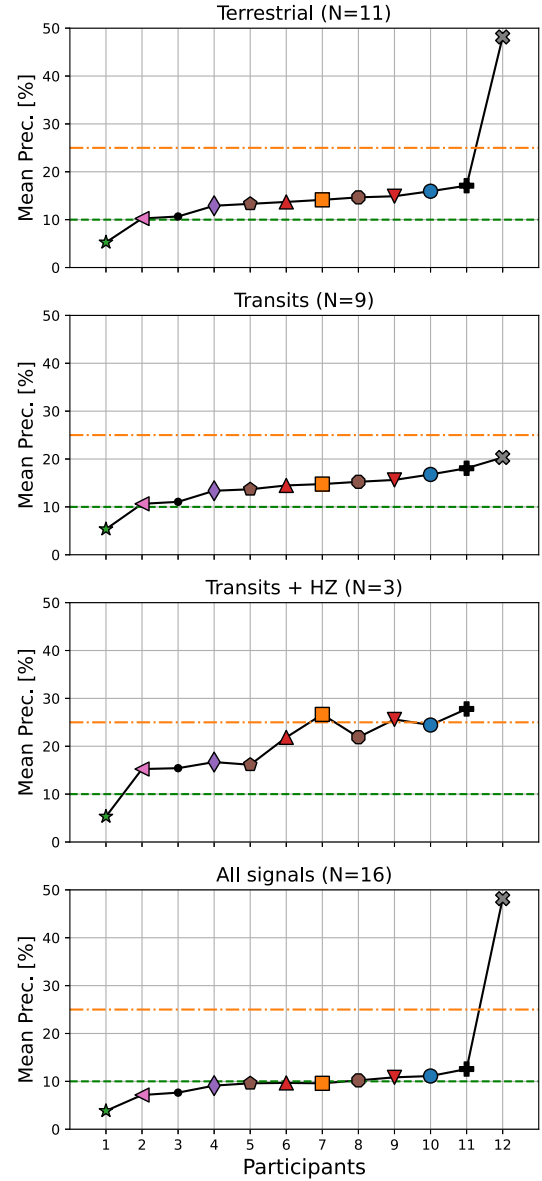


Figure C1. Same as Fig. 10 for the *precision* σ_K instead of *maximum error* ϵ_K . In general it can be seen that the *precision* reached by all the participants is lower than their *accuracy* (Fig. C2), meaning that all the participants tend to slightly overestimate their *precision* in comparison to their *accuracy*. Furthermore, participants with a *precision* better than the white benchmark (black dots ranked third) are certainly underestimating their uncertainties when fitting the Keplerian model.

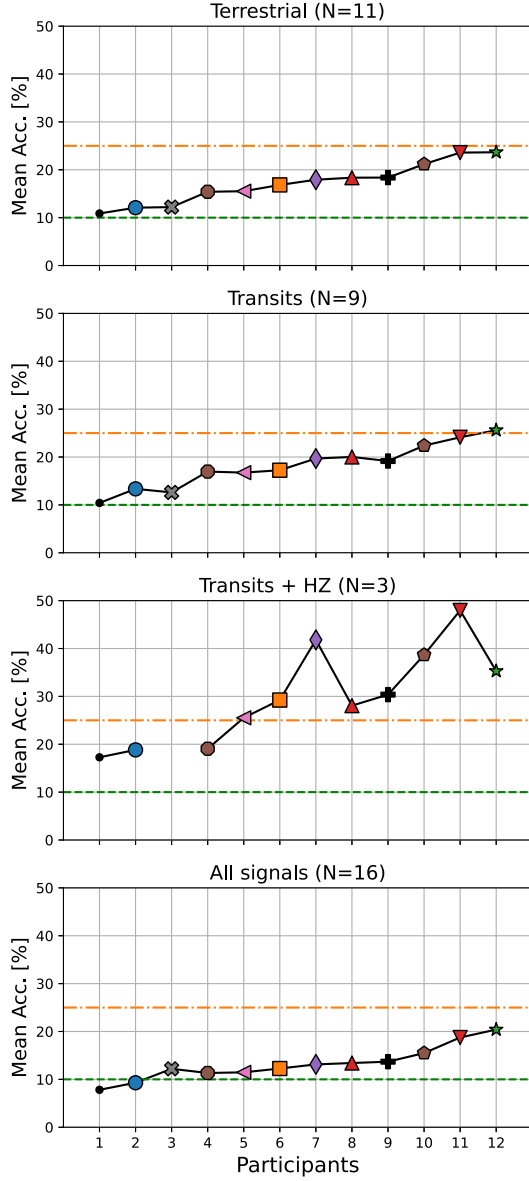


Figure C2. Same as Fig. 10 for the $\text{accuracy} \Delta K/K$ instead of *maximum error* ε_K . The best accuracy reached on the transiting planet by the benchmark was expected to be around 15 per cent.

This paper has been typeset from a $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ file prepared by the author.