

# Exploring the variability of young stars with Gaia DR3 light curves

C. Mas<sup>1,\*</sup>, J. Roquette<sup>1,\*,\*\*</sup>, M. Audard<sup>1</sup>, M. Madarász<sup>2,3,4</sup>, G. Marton<sup>2,3</sup>, D. Hernandez<sup>5</sup>,  
I. Gezer<sup>2,3</sup>, and O. Dionatos<sup>6</sup>

<sup>1</sup> Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland

<sup>2</sup> Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Hungarian Research Network (HUN-REN), Konkoly Thege Miklós Út 15–17, 1121 Budapest, Hungary

<sup>3</sup> CSFK, MTA Centre of Excellence, Budapest, Konkoly Thege Miklós út 15–17, 1121 Budapest, Hungary

<sup>4</sup> Department of Experimental Physics, Institute of Physics, University of Szeged, Dóm tér 9, 6720 Szeged, Hungary

<sup>5</sup> Universität Wien, Institut für Astrophysik, Türkenschanzstrasse 17, 1180 Wien, Austria

<sup>6</sup> Natural History Museum Vienna, Burgring 7, 1010 Vienna, Austria

Received 13 March 2025 / Accepted 20 October 2025

## ABSTRACT

**Context.** Photometric variability is a defining characteristic of young stellar objects (YSOs) that can be traced back to a range of physical processes that occur at different stages in the formation and early evolution of young stars. The Gaia third Data Release (GDR3) has provided an unprecedented dataset of photometric time series, including 79 375 light curves for sources classified as YSO candidates. Through its all-sky coverage, Gaia provides a unique opportunity for large-scale studies of YSO variability.

**Aims.** Our goal was to characterise the GDR3 sample of YSO variables to better identify the recurrence of YSO variability modes (caused by accretion, extinction, rotation modulation, etc.). We made a pilot study of the applicability of the asymmetry (M) and periodicity (Q) variability metrics to characterise YSO variability with Gaia light curves. By adapting the Q–M metrics for sparse and long-term light curves, we sought to bridge the gap between low- and high-cadence survey insights on YSO variability.

**Methods.** We adapted the Q–M method for Gaia. Through a refined sample selection, we identified sources with an appropriate sampling for the Q–M method. We used the generalised Lomb Scargle periodogram and structure functions to derive characteristic variability timescales.

**Results.** We successfully derived Q–M indices for 23 000 sources in the GDR3 YSO sample. These variables were then classified into eight variability morphological classes. We linked the morphological classes with physical mechanisms using  $H\alpha$  as a proxy of accretion and  $\alpha_{\text{IR}}$  indices to gauge whether circumstellar material was present.

**Conclusions.** We demonstrate that the Q–M metrics can be successfully applied to study the sparse time series of Gaia. We applied it successfully to distinguish between the various variability modes of YSOs. While our results are generally consistent with previous high-cadence short-term studies, the long GDR3 time span yields a larger variety of variability mechanisms.

**Key words.** stars: formation – stars: pre-main sequence – stars: protostars – stars: rotation – stars: variables: general – stars: variables: T Tauri, Herbig Ae/Be

## 1. Introduction

Gaia (Gaia Collaboration 2016; Gaia Collaboration 2023b) is a cornerstone mission by ESA that is in operation since mid-2014. In 2022, the Gaia Data Release 3 (GDR3) provided photometric and astrometric observations for 1.8 billion stars. GDR3 included epoch photometry covering 34 months of the mission provided for 11 754 237 variable sources (stellar and extragalactic; Eyer et al. 2023) and all sources in the field of view (FOV) of the Gaia Andromeda Photometric Survey (GAPS Evans et al. 2023). The variables were classified into 25 classes with the GDR3 general variability detection path via supervised machine learning (Rimoldini et al. 2023), and 79 375 of them were classified as young stellar objects (YSOs). This sample of YSO candidates was validated by Marton et al. (2023) by cross-matching it with existing YSO catalogues from optical and infrared surveys. This revealed that 40 000 sources were newly identified YSOs. In this

paper, we further explore the GDR3 YSO variable sample by studying their variability modes in more detail.

Variability is a prevalent characteristic of YSOs, which are observed throughout their entire electromagnetic spectrum (Joy 1945; Herbst et al. 1994; Stelzer et al. 2007; Carpenter et al. 2001; Morales-Calderón et al. 2011; Venuti et al. 2015; Roquette et al. 2020). This variability can be traced back to a wide range of physical mechanisms that occur in the formation and evolution of young stars, including mass accretion, star-disc interaction, disc dissipation, and outflows (e.g. Fischer et al. 2023). Understanding YSO variability is thus fundamental to advancing our knowledge of how stars form and evolve during their first million years.

Attempts to devise variability classification schemes capable of grasping the assortment of observed YSO light curves have been made since the 1990s (e.g. Herbst et al. 1994). Although light curve characteristics such as amplitude, timescale, and periodicity were recurrent ingredients, these earlier classification schemes were limited to the specific data for which they were designed. Based on an initial sample of 162 YSOs in the

\* These authors contributed equally to this work.

\*\* Corresponding author: [julia.roquette@unige.ch](mailto:julia.roquette@unige.ch)

young cluster NGC 2264, [Cody et al. \(2014\)](#) proposed a new morphology classification scheme for YSO light curves. This scheme stems from two variability metrics designed to quantify the level of asymmetry (M index) and periodicity (Q index) in the light curves and allow for sorting sources into seven variability classes: strictly periodic (P), eclipsing binary (EB), quasi-periodic symmetric (QPS), quasi-periodic dippers (QPD), aperiodic dipper (APD), burster (B), and stochastic (S). Two more classes, multi-periodic and long trend (L), were added to include sources with characteristics that were missed by the default Q–M classification. Previous research comparing Q–M classification results with proxies of young stellar evolution (e.g. disc IR excess and accretion-induced H $\alpha$  emission) helped to associate these classes with various physical mechanisms behind the variability of YSOs. A scenario emerged according to which P light curves are attributed to magnetically induced cold spots on the stellar surface viewed from high-inclination angles (e.g. [Cody et al. 2022](#)). Higher-amplitude QPS light curves are associated with accretion-induced hot spots at the stellar surface (e.g. [Venuti et al. 2015](#); [Cody & Hillenbrand 2018](#)). Dipper variability is linked to the AA Tauri phenomenon (e.g. [Bouvier et al. 1999](#); [Alencar et al. 2010](#)), where magnetic star-disc-interaction is thought to lift dusty disc material above the disc mid-plane, which leads to periodic or aperiodic occultation. Bursting variables have been correlated with observation of intense ongoing accretion, suggesting a common origin (e.g. [Stauffer et al. 2014](#)). S variability is thought to originate from accretion-related processes combined with low viewing angles ([Cody & Hillenbrand 2018](#)). While the Q–M classification scheme was initially designed with the aim to apply it to high-cadence space-based light curves from the Convection, Rotation and planetary Transits satellite (CoRoT [Auvergne et al. 2009](#)) and Spitzer telescope, it has since been employed in the context of varied epoch surveys including other high-cadence surveys such as the Kepler mission K2 ([Cody & Hillenbrand 2018](#); [Venuti et al. 2021](#); [Cody et al. 2022](#)), but also ground-based surveys such as the Zwicky Transient Facility (ZTF; [Hillenbrand et al. 2022](#); [Wang et al. 2023](#); [Jiang & Hillenbrand 2024](#)), the All-Sky Automated Survey for Supernovae (ASAS-SN; [Bredall et al. 2020](#)), and the Vista Variables in the Vía Láctea Extended (VVVX; [Ordenes-Huanca et al. 2024](#)).

We present the results of a pilot study that addressed the applicability of the Q–M metrics to GDR3 light curves. It is challenging to apply this method because the observational window resulting from the Gaia scanning law is sparse and unevenly spaced. Most sources have only 30 to 100 epoch observations that span 900 days, which requires a full adaptation of the original method by [Cody et al. \(2014\)](#), based on 40-day-long high-cadence light curves with more than 4000 epochs. Still, Gaia offers all-sky coverage, which is in contrast to the modest sample of a few hundred YSOs that were studied by high-cadence surveys focused on a handful of nearby star-forming regions. Beyond the caveats of its sparse sampling, Gaia therefore provides the opportunity to derive general statistics of YSO variability from a more extensive and diverse sample. Additionally, the longer time coverage of Gaia allowed us to probe longer variability timescales, which are relevant for our understanding of the scope and prevalence of YSO variability mechanisms.

This paper is organised as follows. Sect. 2 describes the data we used. With the goal of testing whether the Q–M indices are applicable to GDR3 data, we performed several tests that we describe in Sect. 3 to identify and discard sources or epoch observations that might be affected by instrumental biases. The clean sample was divided into two variability levels, and the

highest-variability sample was used to benchmark our method. In Sect. 4, we describe our implementation of the Q–M indices for GDR3, with timescales derived using generalised Lomb–Scargle periodograms or structure functions. We also discuss the cases without reliable timescale or those where the timescale is too long compared to the full light curve duration. Sect. 5 presents the results for the Q–M indices, timescales, and morphological classes. We also look into the Q–M indices in relation to the disc properties and evolutionary stage of our samples, and we use the  $\alpha_{\text{IR}}$  index and the H $\alpha$  emission line. In Sect. 6, we discuss our results and present example light curves from each class, and we discuss the typical timescale, amplitude, and physical mechanisms behind them. We further discuss the limitations of our methods and compare our results to previous high-cadence and ground-based studies. Finally, we summarise our findings in Sect. 7.

## 2. Data

### 2.1. Epoch data

GDR3 epoch data cover 34 months of Gaia operation from July 25, 2014 to May 28, 2017. We used two epoch photometry datasets from GDR3:

- YSO sample: composed of 79 375 variable sources identified as YSO candidates as part of the GDR3 variability processing ([Eyer et al. 2023](#); [Rimoldini et al. 2023](#); [Marton et al. 2023](#)).
- GAPS: The Gaia Andromeda Photometric Survey (GAPS [Evans et al. 2023](#)) released epoch photometry for over 1.2 million sources located in a region of 5.5° radius centred on Andromeda. GAPS included the release of epoch photometry for all sources, i.e. both variables and non-variables, showcasing the full power of the Gaia epoch data.

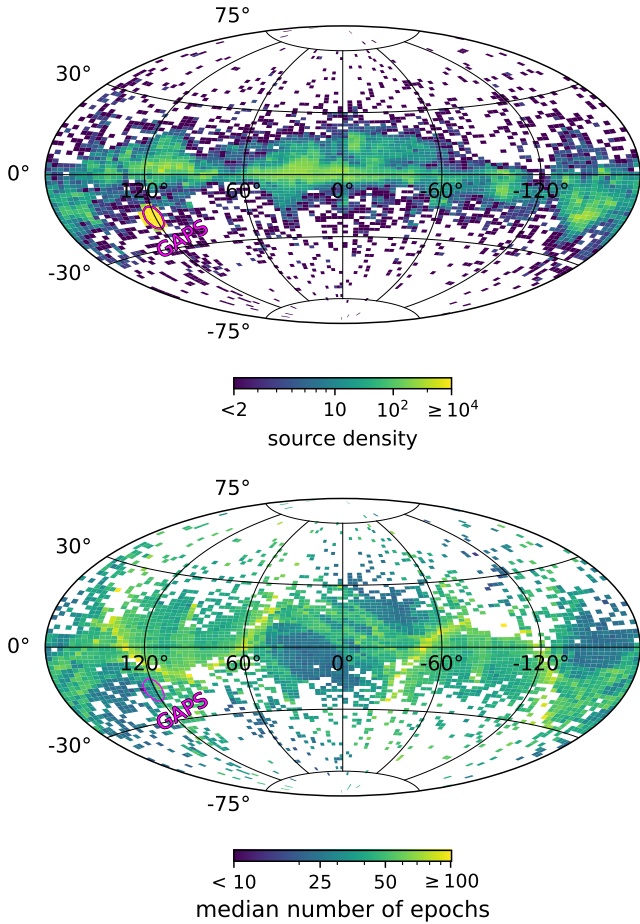
Because of the Gaia scanning law, the number of transits per source highly depends on the sky position. This is illustrated in Fig. 1. For the YSO sample, the number of epochs ranged from 14 to 138, and from 5 to 110 for GAPS. On average, YSO light curves have more epoch observations than GAPS sources. General source properties (mean magnitude, coordinates, etc.) were retrieved from the [Gaia ESA Archive](#). Light curves were retrieved using the `astroquery.gaia` package ([Ginsburg et al. 2019](#)). For this study, we focus our analysis on the G-band epoch data with the goal of classifying variability types. Analysis of other bands ( $G_{\text{BP}}$ ,  $G_{\text{RP}}$ ) may help distinguish between specific YSO processes, but are beyond our current scope.

### 2.2. Supplementary data

In Sect. 3.3, we use data from GDR3 spurious variability table (`vari_spurious_signals`; [Holl et al. 2023](#)) to identify sources affected by scan-angle spurious variability. This data was retrieved directly from the [Gaia ESA Archive](#).

#### 2.2.1. Epoch data from the literature

In Sect. 6, we compare our results with previous literature investigating YSO variability using K2 and ZTF light curves. In [Cody et al. \(2022\)](#), K2 light curves were employed to study sources in the Taurus star-forming region, with 61 sources in common with our YSO sample; light curves were provided to us through private communication with the authors. [Hillenbrand et al. \(2022\)](#), [Wang et al. \(2023\)](#), and [Jiang & Hillenbrand \(2024\)](#) used ZTF light curves to study YSO variability in the North America and



**Fig. 1.** Sky distribution of sources with GDR3 epoch data (YSO and GAPS samples). The location of the GAPS field is highlighted by an ellipse and a label in magenta. The distributions are shown in Galactic coordinates and in an Aitoff projection. In the top panel, colours indicate the source density, and in the bottom, colours reflect the average number of epochs for the light curves given their sky position.

Pelican Nebulae Complex, the Perseus Molecular Cloud, and the Mon R2 region, respectively. These studies included 509 sources in our YSO sample. ZTF light curves for these sources were downloaded from the ZTF archive (Masci et al. 2019) through the ZTF data access tool at the NASA/IPAC Infrared Science Archive. The time span of these light curves was trimmed to match the data release used in these previous studies, and the pre-processing of light curves followed the steps described in these studies.

### 2.2.2. YSO properties

To support validation of our results in Sect. 6, we matched our sources with catalogues that provide measurements related to the youth and evolutionary stages of YSOs. For measurements related to the  $H\alpha$  line, a proxy of accretion, we used:

- The NEMESIS (New Evolutionary Model for Early Stages of Stars with Intelligent Systems) Catalogue of YSOs for the Orion star formation complex (Roquette et al. 2025) compiles youth indicators for sources in that region, including  $H\alpha$  measurements for 5231 objects in our YSO sample.
- GALAH DR4 (Galactic Archaeology with HERMES Survey Data Release 4; Buder et al. 2025) includes  $EW(H\alpha)$  from high-resolution spectroscopy for 2085 of our sources.

- LAMOST DR10 v2.0 (The General Survey of the Large Sky Area Multi-Object Spectroscopic Telescope Data; Zhao et al. 2012) includes  $EW(H\alpha)$  measured from low-resolution spectra for 2172 sources.
- The Gaia-ESO Survey (Gilmore et al. 2022; Randich et al. 2022) DR5 5.1 catalogue (Hourihane et al. 2023) provides  $EW(H\alpha)$  and full widths at 10% of the peak flux measurements from high-resolution spectra for 751 and 765 sources, respectively.
- Gaia DR3 includes spectroscopic data products based on low-resolution spectra from the blue and red photometers (De Angeli et al. 2023), providing measurements of  $EW(H\alpha)$  with pseudo-equivalent widths estimated as part of the Extended Stellar Parametriser for Emission-Line Stars (ESP-ELS Creevey et al. 2023) for 56 964 sources.

## 3. Data pre-processing

Fig. 2 summarises the pre-processing and filtering steps we employed to ensure that only the most reliable light curves were kept in our sample. Our first cleaning step was to remove likely spurious epoch observations based on GDR3 flag `rejected_by_variability`, which provides information on transits considered as outliers or extreme values and rejected as part of GDR3 variability processing steps (see Sect. 3.1 of Eyer et al. 2023).

### 3.1. Contamination by extragalactic sources

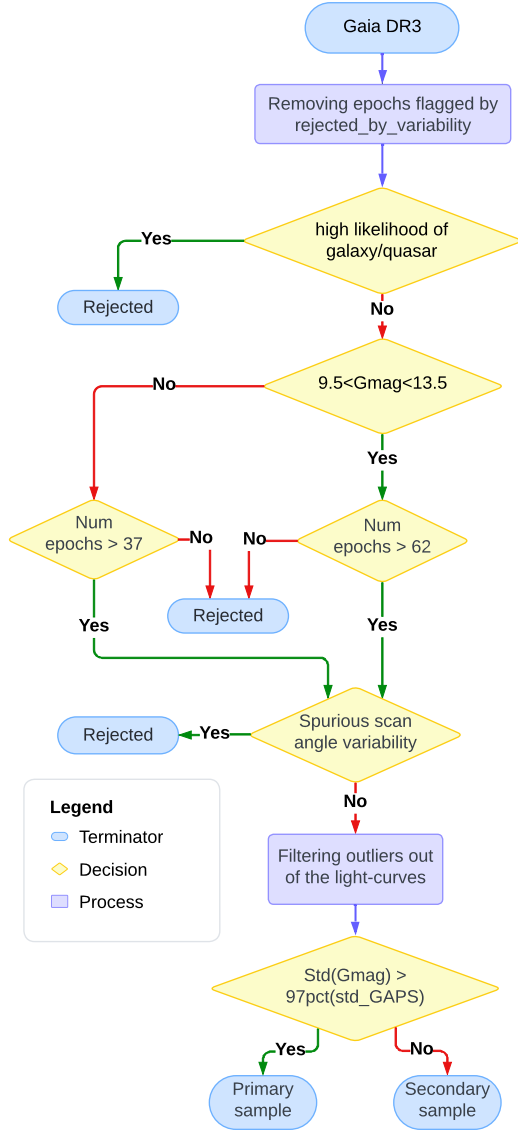
We excluded 14 722 sources from the GAPS sample which were flagged with a probability over 80% of being a galaxy or a quasar, based on the Gaia DR3 classification probabilities `classprob_dsc_combmod_galaxy` and `classprob_dsc_combmod_quasar` (Gaia Collaboration 2023a; Delchambre et al. 2023). Six sources were similarly removed from the YSO sample.

### 3.2. Minimum number of epochs

In Appendix A.1, we performed tests to determine the minimum number of epochs required in the light curve to ensure statistically robust variability indices. All tests were carried out as a function of magnitude to account for the different levels of photometric noise in the sample. Based on these tests, we rejected light curves with fewer than 62 observations for  $G = 9.5$ – $13.5$  mag and 37 for outside this range. This filtering step yield 44 918 sources in our YSO sample (57% of the full catalogue) and 524 096 sources in the GAPS sample (42%).

### 3.3. Spurious variability

Holl et al. (2023) investigated spurious variability in GDR3 light curves due to scan-angle effects (combination of the Gaia mirror aspect ratio, spin-axis precession, and orbital period, Gaia Collaboration 2016). Attempts to model and filter out the spurious frequencies caused by these scan-angle effects would require assumptions such as the underlying variability signal having a sinusoidal waveform, which are not reliable priors for YSO variability. Hence, instead of attempting to mask these signals, we followed the Gaia data-processing literature (Holl et al. 2023; Distefano et al. 2023) to identify and remove sources likely affected by spurious scan-angle signals (see Appendix A.2). After this step, we deemed 34 793 light curves in the YSO



**Fig. 2.** Workflow of filtering and cleaning steps to remove extragalactic contamination prior to asymmetry and periodicity analysis (Sect. 3.1), sources with too few epochs (Sect. 3.2), affected by a spurious scan-angle (Sect. 3.3), outliers (Sect. 3.4), and, finally, sorting the sources based on their degree of variability (Sect. 3.5).

sample as reliable for further analysis (44% of the full catalogue), and 479 936 in the GAPS sample (39%).

### 3.4. Outlier removal

Filtering large outliers is an important step for studying the degree of asymmetry in the light curves, as outliers can bias the M index calculation (see Sect. 4.1). When dealing with high-cadence data, Cody et al. (2014) addressed this issue by  $5\sigma$ -clipping the light curves. With the densely sampled light curves of CoRoT and K2, this approach was unlikely to remove relevant variability features. However, due to the sparsity of the GDR3 light curves, this type of filtering needs to be addressed more carefully. We have done so in Appendix A.4, where 636 sources in our sample had at least one outlier removed. All affected sources still fulfilled the minimum number of epochs requirement after outlier removal.

### 3.5. Selection of highly variable sources

Initial tests revealed that low-amplitude QPS was the most common variability mode in the YSO sample, dominating the light curves of 41% of the sources analysed (see Sect. 6.2.2 and Table 1). However, to assess the applicability of the Q–M indices to GDR3 light curves, we concentrated our pipeline development on a sample selected to include a higher proportion of less common variability modes. This was achieved by focussing our initial analysis on a sub-sample of the most variable sources.

To support the identification of the most variable sources as a function of magnitude, it was relevant to first understand the behaviour of the YSO sample compared to non-variable stars also observed by GDR3. Although GDR3 does not provide light curves for non-variable YSOs, this comparison was still possible based on the GAPS sample, which includes GDR3 light curves for all sources in the Andromeda field, irrespective of their variability status. The GAPS and YSO samples share GDR3 photometric performance, albeit having different sky locations and numbers of transits (see Fig. 1). A comparison between the two samples as a function of their average G magnitude and standard deviation (std) is provided in Fig. 3, where we display properties of light curves that passed pre-processing steps described in this section (34 793 YSO and 479 334 GAPS light curves), i.e. already accounting for the minimum number of epochs required for stable variability indices (Sect. 3.2, Appendix A.1). The GAPS sample is predominantly made up of non-variables, with roughly 1% of its sources selected by GDR3 variability processing. This is evident in Fig. 3, where the bulk of the GAPS distribution (yellow regions, top panel) is located below the lower envelope of the YSO sample distribution (derived as the 1st rolling percentile of that distribution).

We defined a threshold for selecting the most significant variables as a function of magnitude based on rolling percentiles of the GAPS distribution, which were estimated by binning the GAPS distribution using 0.05 mag bins from  $G = 9.1$ – $21.9$  mag. Each magnitude bin was required to contain at least 500 sources, and bin sizes were sometimes increased by a factor of 1.1 until this minimum number of sources was satisfied. Rolling percentiles were then calculated within each bin, and a Savitzky–Golay filter (Savitzky & Golay 1964) was applied to smooth it. The 97<sup>th</sup>-percentile threshold provided the best compromise between selecting highly variable sources while still retaining a large enough sample of variables.

Finally, we selected 10 739 sources in the YSO sample with a std larger than 97% of GAPS sources in their magnitude bin. These sources are the primary sample on which we focused the adaptation of the Q–M method (Sect. 4). The remaining 24 054 sources below the 97<sup>th</sup> percentile of the GAPS distribution are hereafter named secondary sample; these were also analysed once the workflow for the Q–M analysis was settled.

## 4. Methods

### 4.1. Asymmetry and periodicity metrics

We explored the applicability of the Q–M indices to characterise the GDR3 light curves. The Q index is defined as (Cody et al. 2014; Hillenbrand et al. 2022):

$$Q = \frac{\text{std}_r^2 - \langle \sigma^2 \rangle}{\text{std}_m^2 - \langle \sigma^2 \rangle}, \quad (1)$$

where  $\text{std}_r^2$  is the variance of the residuals of the signal, estimated by subtracting a waveform estimation (derived by smoothing the

**Table 1.** Counts and frequencies of sources for each variability morphologic class.

Class	Count	Thick disc	Frequency <sup>a</sup>				Spectral type <sup>b</sup>			
			general disc	CTTS	ts-ratio	ptp	M	K	G	AB-F
Full sample										
P	286	14.3 (2.1)	50.0 (3.0)	13.7 (2.8)	13.4	0.19	77.3 (2.5)	12.9 (2.0)	1.7 (0.8)	8.0 (1.6)
QPS	9514	20.0 (0.4)	57.6 (0.5)	14.8 (0.6)	2.6	0.15	71.2 (0.5)	15.3 (0.4)	4.0 (0.2)	9.5 (0.4)
QPD	5439	39.1 (0.7)	67.3 (0.6)	24.7 (1.1)	1.8	0.35	65.4 (0.6)	18.6 (0.5)	5.4 (0.3)	10.6 (0.4)
APD	1103	56.6 (1.5)	79.9 (1.2)	39.7 (3.7)	0.4	0.53	66.2 (1.4)	19.7 (1.2)	5.7 (0.7)	8.4 (0.8)
B	5252	21.0 (0.6)	56.7 (0.7)	22.0 (1.1)	2.0	0.16	59.8 (0.7)	24.3 (0.6)	3.2 (0.2)	12.7 (0.5)
S	1490	27.9 (1.2)	65.9 (1.2)	23.6 (2.6)	0.8	0.19	71.4 (1.2)	17.8 (1.0)	3.2 (0.5)	7.6 (0.7)
EB	129	15.5 (3.2)	56.6 (4.4)	11.4 (3.6)	23.6	0.20	85.3 (3.1)	9.3 (2.6)	0.8 (0.8)	4.6 (1.9)
L	4966	21.1 (0.6)	48.5 (0.7)	11.0 (1.1)	–	0.16	44.3 (0.7)	22.9 (0.6)	7.7 (0.4)	25.1 (0.6)
Primary sample										
P	199	17.1 (2.7)	46.2 (3.5)	14.6 (3.1)	26.6	0.22	81.4 (2.8)	8.6 (2.0)	2.0 (1.0)	8.0 (1.9)
QPS	2362	41.9 (1.0)	65.5 (1.0)	36.6 (1.6)	1.4	0.31	60.5 (1.0)	22.2 (0.9)	7.1 (0.5)	10.2 (0.6)
QPD	2886	61.7 (0.9)	81.4 (0.7)	40.6 (1.6)	1.3	0.57	65.5 (0.9)	20.5 (0.8)	6.1 (0.4)	7.9 (0.5)
APD	822	68.0 (1.6)	85.2 (1.2)	48.2 (4.2)	0.4	0.67	63.4 (1.7)	21.9 (1.5)	7.0 (0.9)	7.7 (0.9)
B	1505	44.8 (1.3)	69.4 (1.2)	45.6 (2.2)	1.3	0.31	68.4 (1.2)	18.2 (1.0)	4.3 (0.5)	9.1 (0.7)
S	601	42.8 (2.0)	70.9 (1.9)	49.0 (4.9)	0.4	0.30	62.1 (2.0)	23.3 (1.7)	4.8 (0.9)	9.8 (1.2)
EB	95	15.8 (3.7)	51.6 (5.1)	14.1 (4.3)	91.0	0.24	84.2 (3.7)	8.4 (2.8)	1.1 (1.1)	6.3 (2.5)
L	1354	40.0 (1.3)	63.1 (1.3)	33.6 (3.2)	–	0.37	48.6 (1.4)	26.3 (1.2)	8.2 (0.7)	16.9 (1.0)
Secondary sample										
P	87	8.0 (2.9)	58.6 (5.3)	8.7 (5.9)	5.8	0.10	67.8 (5.0)	23.0 (4.5)	1.1 (1.1)	8.0 (2.9)
QPS	7152	12.8 (0.4)	55.0 (0.6)	6.1 (0.5)	3.3	0.09	74.7 (0.5)	13.1 (0.4)	3.0 (0.2)	9.2 (0.3)
QPD	2553	13.4 (0.7)	51.5 (1.0)	4.5 (0.8)	2.5	0.09	65.2 (0.9)	16.5 (0.7)	4.6 (0.4)	13.7 (0.7)
APD	281	23.1 (2.5)	64.4 (2.9)	3.0 (3.0)	0.7	0.13	74.4 (2.6)	13.5 (2.0)	1.8 (0.8)	10.3 (1.8)
B	3747	11.4 (0.5)	51.6 (0.8)	7.0 (0.9)	2.4	0.10	56.4 (0.8)	26.8 (0.7)	2.6 (0.3)	14.2 (0.6)
S	889	17.8 (1.3)	62.5 (1.6)	7.4 (2.0)	1.1	0.12	77.8 (1.4)	14.0 (1.2)	2.1 (0.5)	6.1 (0.8)
EB	34	14.7 (6.1)	70.6 (7.8)	–	6.8	0.10	88.2 (5.5)	11.8 (5.5)	–	–
L	3612	14.0 (0.6)	43.0 (0.8)	2.5 (0.7)	–	0.09	42.6 (0.8)	21.6 (0.7)	7.6 (0.4)	28.2 (0.7)

**Notes.** Morphologic classes defined as in Sect. 4.4. The primary and secondary samples are detailed in Sect. 3.5. Thick-disc frequency indicates the frequency of stars with an envelope or a thick disc, as indicated by the  $\alpha_{\text{IR}} \geq -1.6$ . General disc frequency refers to the frequency of sources with  $\alpha_{\text{IR}} \geq -2.5$  or with EW(H $\alpha$ ) indicative of ongoing accretion. CTTS frequency indicates the frequency of sources identified as CTTS in Sect. 5.3.1 in comparison with the total number of sources with H $\alpha$  information. ts-ratio indicates the ratio of sources with rotation against intermediate timescales. ptp indicates the average peak-to-peak amplitude per class. The primary and secondary sample are detailed in Sect. 3.5. <sup>(a)</sup> Values in parentheses are standard error estimates for the disc frequencies. <sup>(b)</sup> Frequency of spectral types as a function of variability class.

phase-folded light curve, folded to the characteristic variability timescale) from the raw light curve.  $\text{std}_m^2$  is the variance of the original light curve, and  $\langle \sigma^2 \rangle$  is the square of the typical uncertainty (calculated here as the average of all squared errors associated with the data). The Q index is thus a measure of the ratio of the variance of the residuals to the general variance of the light curve. Lower Q values close to zero indicate high periodicity. In contrast, higher Q values (close to unity) reflect irregular or stochastic variability. The derivation of the Q index assumes that the light curve contains a signal that may repeat at a characteristic timescale. This signal can be modelled by phase-folding the light curve to its characteristic timescale and employing a smoothing technique to model the average waveform of the signal. A residual curve is then calculated as the difference between the raw-light curve and this waveform.

The M index measures the asymmetry of the light curve by comparing how the upper and lower brightness deciles deviate from the median. It is defined as

$$M = \frac{\langle m_{10\%} \rangle - m_{50\%}}{\sigma_m}, \quad (2)$$

where  $\langle m_{10\%} \rangle$  is the mean of the highest and lowest 10% of light-curve points,  $m_{50\%}$  is the median magnitude, and  $\sigma_m$  is the

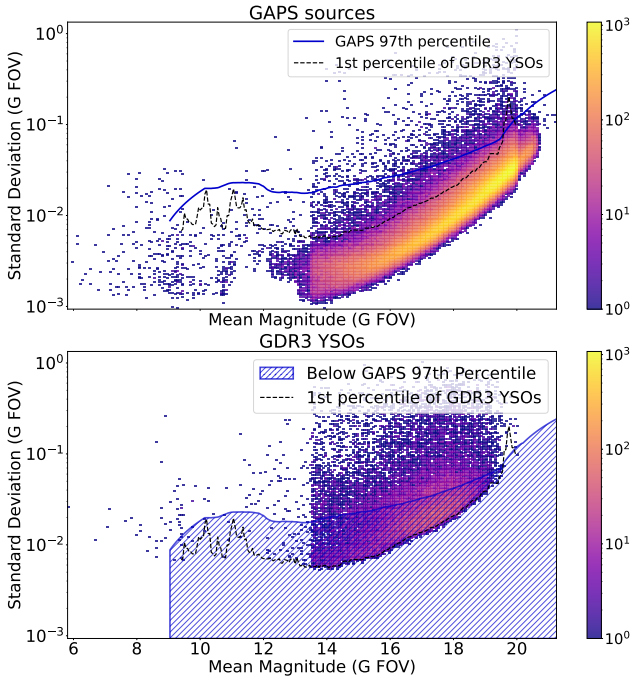
standard deviation. When working in magnitudes, light curves with dipping behaviour have  $M \geq 0.25$ , those with bursts have  $M \leq -0.25$ , and symmetric light curves have  $M \approx 0$ .

## 4.2. Characteristic timescale

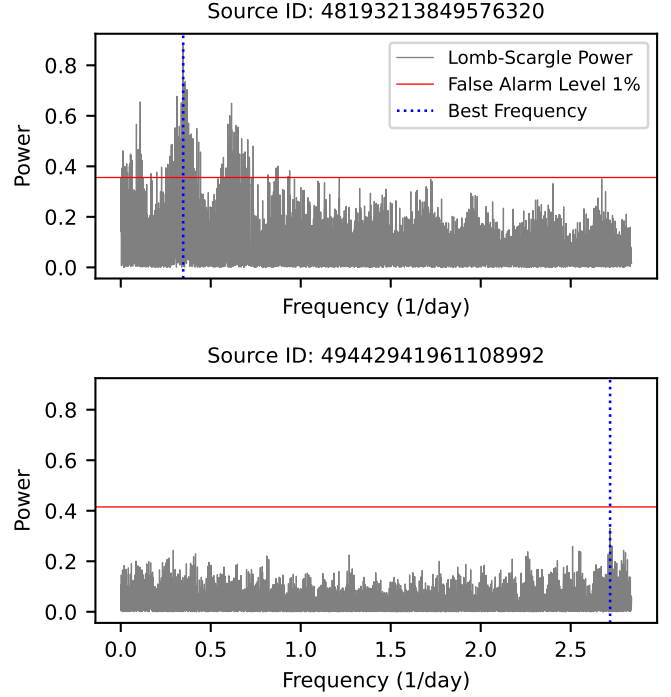
The capital step to quantify the periodicity with the Q index is to derive a characteristic timescale of variability. Fig. 4 summarises our workflow for timescale derivation: We first applied the generalised Lomb–Scargle periodogram to identify periodic timescales, assessing their reliability using the false alarm probability (Sect. 4.2.1). If no reliable periodic timescale was found, we then computed their structure function to derive an aperiodic timescale (Sect. 4.2.2). Sources were discarded if neither method provided a trustworthy timescale. A final step removed sources with timescales affected by the typical spurious variability of Gaia (Sect. 4.3).

### 4.2.1. Generalised Lomb–Scargle periodogram

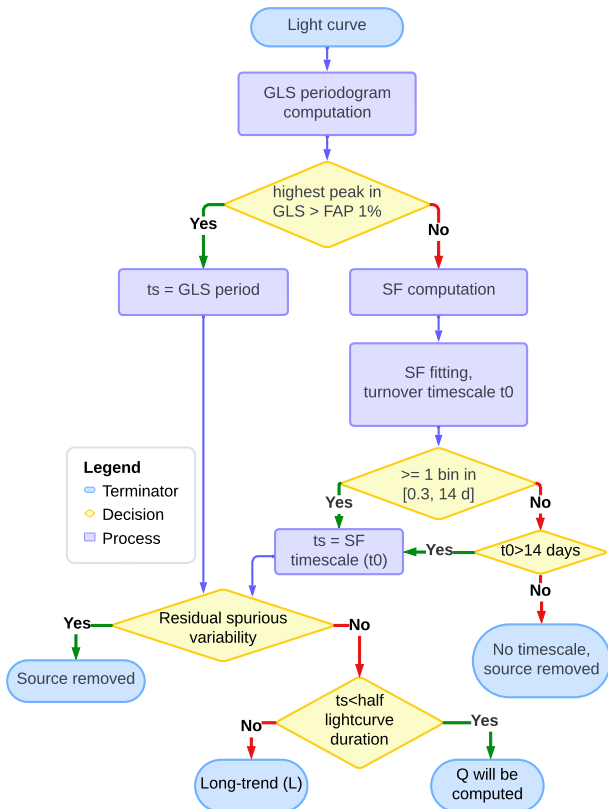
We used the generalised Lomb–Scargle periodogram (GLS; Lomb 1976; Scargle 1982; Zechmeister & Kürster 2009;



**Fig. 3.** Standard deviation versus  $G$  magnitude for light curves in the GAPS (top) and YSO (bottom) samples. The colour bar indicates source counts. The solid blue line marks the 97th percentile for the GAPS sample, and the hatched region identifies YSO sources lying below this percentile (secondary sample), while primary-sample sources lie above it. In both panels, the black dashed line shows the lower envelope (1st rolling percentile) of the YSO distribution to aid visual comparison.



**Fig. 5.** Example GLS periodograms used to derive periodic timescales. The upper panel shows a periodogram with a peak above the 1% FAL, from which a periodic timescale is obtained, while the lower panel shows a case with no such peak, where an aperiodic timescale is instead derived from the structure function.



**Fig. 4.** Workflow for deriving the characteristic timescale required for the periodicity analysis with the  $Q$  index.

(VanderPlas 2018) to search for periodic variability in our sample. We searched for periods between twice the median time step of the light curve (0.5 days) and the entire light curve duration (900 days), with the frequency of the highest peak recorded as the most representative timescale of the light curve variability. The reliability of GLS timescales was evaluated considering a false-alarm probability level (FAL) of 1%, with a Baluev (2008) FAL estimation. Sources without a peak above the FAL were considered aperiodic. Fig. 5 shows examples of GLS periodograms for a periodic and an aperiodic source. A total of 24 067 sources (~69% of the YSO sample) had periodic timescales measured with the GLS. 3721 of these timescales were longer than half the total time span of the light curve, and were stored as evidence of L type variability, but were not used to derive the  $Q$  index.

#### 4.2.2. Structure function

In the previous section, 10 726 sources did not show any peak above the FAL in their GLS, indicative of periodicity. These sources are likely aperiodic variables. Still, their study with the  $Q$  metric required the inference of a characteristic timescale for their aperiodic variability. In their original implementation, Cody et al. (2014) employed the PeakFind (Findeisen et al. 2015) algorithm for aperiodic timescale inference, which requires high-cadence sampling to resolve detailed inflexions in light curves (Cody et al. 2022) and thus is unattainable with sparsely sampled data sets such as GDR3. Alternatively, we inferred aperiodic timescales using structure functions (SF). SFs are widely used in other fields to constrain variability timescales, with varied applications in the study of Active Galactic Nucleus (Simonetti et al. 1985; Kozłowski 2016, AGN), other variable stars (Eyer & Genton 1999), and with a few applications to YSOs

albeit in the context of high-cadence surveys (Venuti et al. 2021; Lakeland & Naylor 2022). In the context of Gaia DR2, structure functions were successfully used to investigate short timescale variables (Roelens et al. 2017).

Following Hughes et al. (1992) and Roelens et al. (2017), the SF of a light curve is defined as

$$SF(\tau) = \frac{1}{N(\tau)} \sum_{i>j} (m_i - m_j)^2, \quad (3)$$

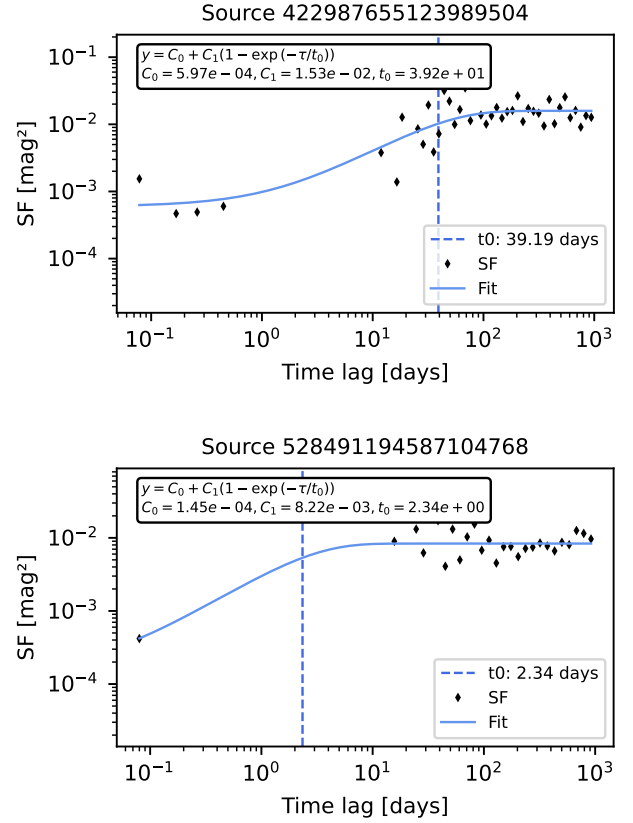
where  $N(\tau)$  is the number of pairs separated by a time difference  $\tau$ , and  $m_i - m_j$  the magnitude difference for each pair. We discretised Eq. (3) using logarithmically spaced bins. Each bin corresponds to a range of time differences ( $\tau_1, \tau_2$ ), and is represented by its midpoint  $\tau$ . We selected all pairs of observations whose time separation  $\Delta t = t_i - t_j$  lies within this range to compute  $SF(\tau)$  at each bin. Bins that do not contain any pairs are excluded from the analysis. To further adapt this binning strategy to unevenly sampled light curves, we introduced a tolerance factor  $\epsilon$ , which ensures that each bin contains enough pairs for a reliable calculation (Roelens et al. 2017). If a populated bin initially has too few pairs, its size is increased symmetrically until the minimum requirement of at least three pairs is met. With GDR3 data, we could compute SFs for time differences in the range 0.07–980 d (see Fig. B.1). The number of bins used was based on the number of epochs available for each light curve, with 1.5 times the number of epochs used for the number of bins.

For aperiodic variability, the SF typically shows three main regimes. A noise-dominated regime where the SF remains roughly constant around the intrinsic noise level in the data. A variability regime where the SF value increases until it peaks at the turnover timescale  $t_0$ . And a plateau where, beyond  $t_0$ , the SF flattens as no new variability occurs. To represent these three regimes, we fitted the SFs with a function of the form

$$y = C_0 + C_1(1 - \exp(-\tau/t_0)), \quad (4)$$

where  $C_0$ ,  $C_1$ , and  $t_0$  are the fitting parameters and  $\tau$  is the time difference. In this model,  $C_0$  is related to intrinsic variability due to the uncertainty of the measurements and  $C_1$  is associated with the real signal variability amplitude. The final fitting parameter  $t_0$  is the turnover timescale, which marks the transition between the variability-dominated regime and the plateau and can be interpreted as the variability timescale in which we are interested.

We fit Eq. (4) to the SF using the MIGRAD routine (James & Roos 1975) implemented in Python as part of the iminuit library (Dembinski et al. 2020). The initial guesses and limit values for the model parameters were based on inspection of the distribution of all SF shown in Fig. B.1 and were chosen as follows:  $C_0 = 0.02$ ,  $C_1 = 0.08$ ,  $t_0 = 15$  d for the initial guess. While  $C_0$  is in the range  $(10^{-5}, 0.07)$  mag<sup>2</sup>,  $C_1$  in  $(10^{-3}, 5.5)$  mag<sup>2</sup> and  $t_0$  in  $(0.07, 980)$  d. Examples of SF model fit results can be seen in Fig. 6. The imposition of model parameter limits resulted in a relative excess of fit results with edge values. Hence, we filtered these results to remove spurious models within 5% of the edge values for  $C_0$ ,  $C_1$ , and for the lower limit of  $t_0$ . A total of 5354 SF fits were discarded at this step. Visual inspection of discarded models showed that these were typically sources with complex multi-mode variability, including different timescales and that could not be represented by Equation (4), or sources with poorly sampled SF with a long time range without data (blind time range). An excess of upper limit values of  $t_0$  also existed, and we verified that these were typically related to light



**Fig. 6.** Examples of sources with aperiodic timescale derived from fitting an aperiodic model to their binned SF. The best-fit models are shown in blue. The top panel shows an aperiodic variable with a timescale of 39 days. The lower panel presents an example of a source for which no reliable timescale could be determined, because Gaia’s sparse sampling produces a blind interval in its SF.

curves exhibiting a long-term trend (e.g. a gradual change in brightness) with a timescale longer than the time span of GDR3 observations, where the SF continues to increase without ever reaching the plateau, in which case, we stored the timescale as evidence for long-term variability but did not use them to evaluate the Q index. In addition to the issue of edge values, 12 best-fit models were also discarded because they had limited  $\tau$  sampling around the derived timescale. This issue is further discussed in the Appendix B, and it is introduced by GDR3 sparse time-series sampling, which for some sources results in a blind time range in the SF. For affected sources (see Fig. 6-bottom), no data points exist in the range  $\tau \sim 0.3$ –28 days. Beyond this range, the SF show a flat distribution – hence, it is impossible for models to correctly identify the transition from the variability regime into the plateau.

The fits of the SF model in this section allowed us to estimate the aperiodic variability timescale for 4118 sources. For 1245 sources, the SF-fit resulted in upper-edge values or timescales that were longer than half of the total time span of the light curve, with these stored as evidence of long trend, but were not used to evaluate the Q index. For 5363, we were unable to derive a time scale of variability using the GLS or SF methods.

#### 4.3. Residual spurious variability

We further discarded 1251 sources as their measured timescale coincided with typical timescales caused by the spurious variability introduced by the Gaia scan-angle effects (Sect. 3.3, see

also Appendix A.4). After this final cleaning step, we proceeded with the analysis of 28 179 YSO light curves, with 9824 of these in the primary sample and 18 355 in the secondary sample.

#### 4.4. Q and M implementation and morphology classes

Once a characteristic timescale was derived, light curves were folded to their corresponding phase, and a waveform was derived using a Savitzky–Golay filter with a window size of 25% of the phase and a second-order polynomial. This set of parameters allowed us to obtain waveforms representative of the ongoing variability while avoiding overfitting. The estimated waveform was then subtracted from the data to obtain the residual curve, followed by estimating the Q indices with Eq. (1).

Following Cody et al. (2014) scheme, the Q and M indices can be used to gauge light curve morphologies and sort variables into seven standard classes:

- EB:  $Q < 0.11$  and  $M > 0.25$ .
- P:  $Q < 0.11$  and  $M \leq 0.25$ .
- QPS:  $0.11 \leq Q \leq 0.61$  and  $|M| \leq 0.25$ .
- QPD:  $0.11 \leq Q \leq 0.61$  and  $M > 0.25$ .
- APD:  $Q > 0.61$  and  $M > 0.25$ .
- B:  $Q > 0.11$  and  $M < -0.25$ .
- S:  $Q > 0.61$  and  $|M| \leq 0.25$ .

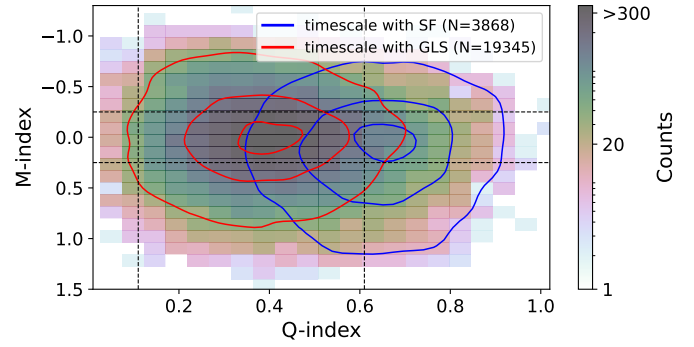
Additional classes are sometimes included in the literature, such as multi-periodic, non-variable, and L. These classes are not based directly on the Q and M indices, but are defined based on the failure to apply Q–M method. The identification of multi-periodic variability in previous studies relies on the high cadence of their observations and requires additional steps, including visual inspection of the light curves. We did not reproduce those steps as part of this study. As our sample is focused on light curves with enhanced variability, the non-variable YSOs are also excluded. L variables were identified here as sources with inferred timescales longer than half of GDR3 time span.

## 5. Results

### 5.1. Asymmetry and periodicity indices

We successfully derived Q indices for 23 213 sources (8385 in the primary sample) and M indices for 28 179 sources (9824; primary). Our successful adaptation of the Q–M method for use with GDR3 light curves is illustrated in Fig. 8 and Fig. 9. In Fig. 8, light curves at low Q values are very well represented by the estimated periodic waveform. At intermediate Q values, quasi-periodic waveforms are still found, but with increasing phase dispersion. At  $Q \approx 1$ , light curves are dominated by stochastic variability, timescales inferred can no longer be interpreted as a periodic timescales, with data points stochastically distributed around the average. In Fig. 9, light curves with extreme asymmetries have M values farthest away from 0, with negative M values corresponding to variables with brightening events and positive M values corresponding to dimming events, while more symmetric light curves are closest to  $M \approx 0$ .

Fig. 7 shows the distribution of Q versus M for our sample. This general distribution peaks at  $Q \sim 0.43$  (quasi-periodic) and  $M \sim 0.05$  (symmetric). The individual distributions of the Q and M indices are shown in Fig. 10, where we separate our sample by timescale method. As expected from the nature of the methodology for measuring variability timescales, the distribution of Q for sources with a periodic timescale derived with the GLS (in red) is mainly on the periodic and quasi-periodic regions of the diagram, peaking at  $Q \sim 0.38$ , with the Q index



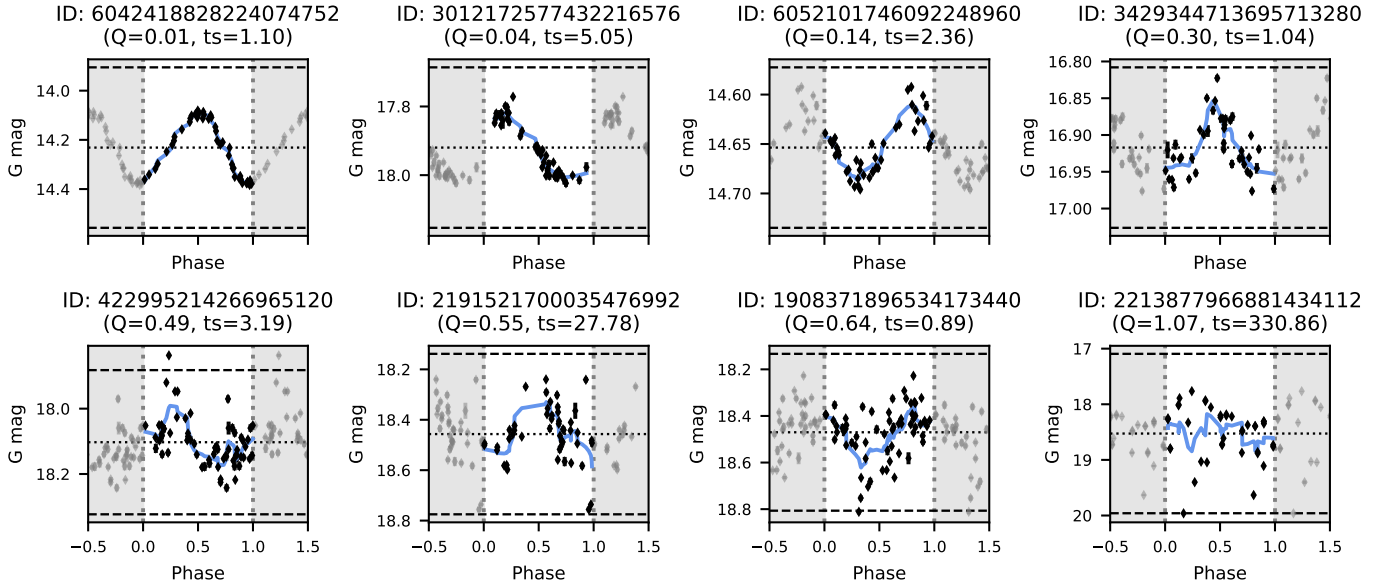
**Fig. 7.** Distribution of Q and M indices the GDR3 YSO sample. Red and blue lines contour the distribution of sources with GLS periodic timescale, and SF aperiodic timescale respectively (Sect. 4.2). The dashed lines show the adopted thresholds for Q–M classification. The three contour curves contain 10, 50, and 90% of the data points, respectively.

distribution for sources with aperiodic timescales derived from SFs peaking at  $Q \sim 0.64$ . For the M index, sources with periodic timescales peak at  $\sim 0.02$ , and at  $\sim 0.15$  for aperiodic timescales, with the aperiodic distribution showing a tail towards dipping M index values. Fig. 10 also compares the distribution of Q–M indices for our primary and secondary samples, where the differences for these two are also supported by Kolmogorov–Smirnov (KS) tests. Quasi-periodic variability is the most frequent mode in our sample, accounting for more than 69% of sources. The secondary sample peaks at  $Q \sim 0.38$  and  $M \sim 0.02$ , being dominated by QPS sources. The over-representation of low-amplitude QPS sources in the general occurrence of variability classes motivated discussing results as a function of higher-amplitude (primary) vs lower-amplitude (secondary) variable samples. The primary sample of variables peaks at  $Q \sim 0.44$  and  $M \sim 0.18$ , with a larger proportion of dipping and aperiodic variables in comparison to the secondary sample.

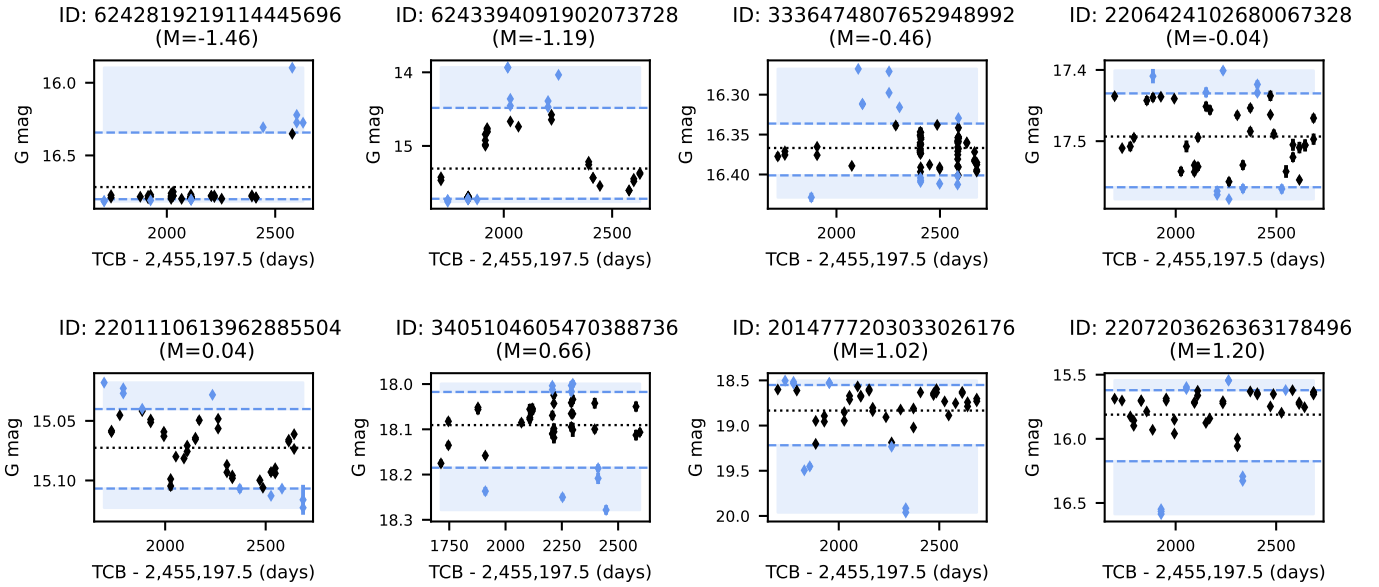
The classification scheme outlined in Sect. 4.4 was applied to sort light curves into seven morphological classes with an eighth class attributed to account for L variables, which had characteristic timescales longer than half the total time span of the GDR3 observations (typically  $\sim 475$  days). The count and frequency of sources in each class are summarised in Table 1. As already suggested by the Q–M distributions, dipper variability is the most common among the primary sample, with  $\sim 38\%$  of these sources showing APD or QPD variability. We identified 4966 L variables, 3721 with periodic timescales derived with the GLS and 1275 aperiodic timescales derived with the SF. We visually inspected a large fraction of these light curves and confirmed that a long-term trend was indeed present. Their timescales are at the limit of GDR3 resolvable timescales and should be interpreted as a bottom limit for the real timescale, and hence we did not employ them to calculate the Q index, although we discuss them as part of our results.

### 5.2. Amplitude and timescales

We further complemented our results analysis with peak-to-peak (ptp) amplitudes and timescales measurements. We used the min-max amplitude as a measure of the ptp amplitude rather than the 5–95 percentile amplitude more commonly employed in the literature, as we found the latter to underestimate ptp amplitudes when applied to GDR3 sparse sampling. Throughout the discussion of results, we refer to timescales in the range



**Fig. 8.** Phase-folded GDR3 light curves are shown with  $Q$  values increasing from left to right. Black diamonds indicate the phase-folded epoch photometry, while the blue line represents the inferred waveform. The grey region marks the extended phase for clarity. Grey horizontal lines denote the mean (dotted) and the mean  $\pm 3 \times$  std (dashed).

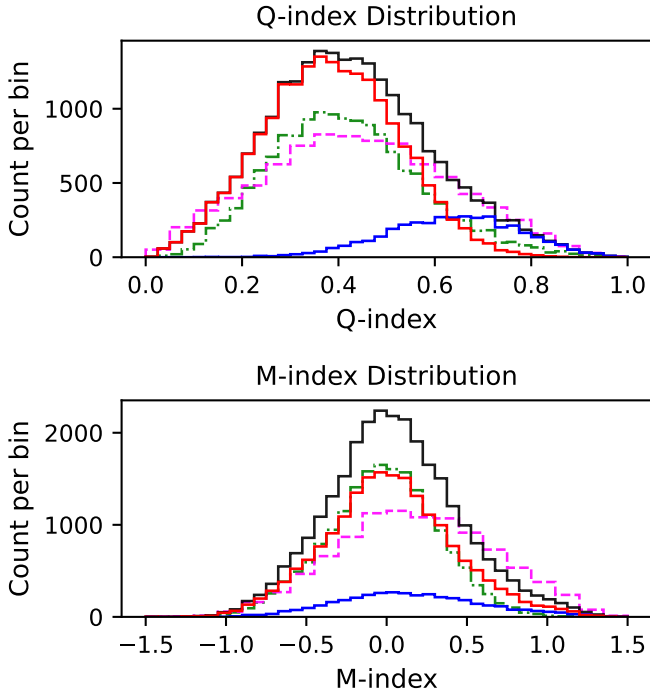


**Fig. 9.** Examples of GDR3 light curves with increasing  $M$  values from left to right. The blue diamonds are the data points of the top and bottom deciles of the curve; the grey dotted line shows the mean; the blue dashed lines show the top and bottom deciles. The time axis reflects the observations' Barycentric Coordinate Time (TCB).

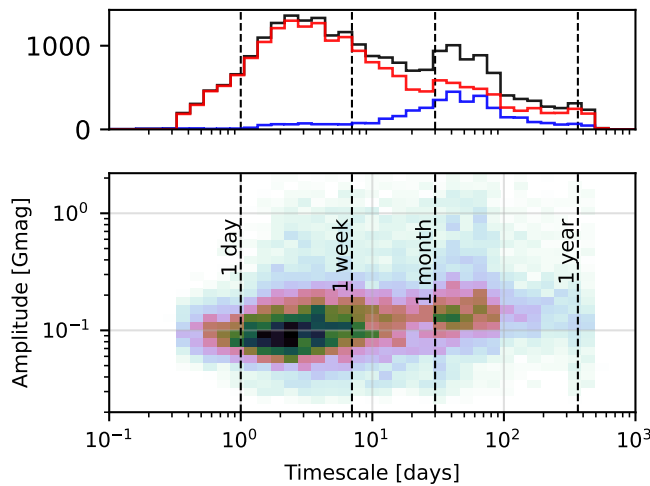
0.5–16 d as “rotation timescales”. Timescales larger than 20 d and less than about half of the duration ( $\sim 475$  days) of the observations are termed “intermediate timescales”. Timescales longer than half of the duration of the light curve are referred to as “long-trend timescales”. Fig. 11 shows the distribution of amplitudes as a function of timescales for our sample, revealing a bimodal distribution of timescales with a population of sources with rotation-timescales (62%) and a second population of variables with intermediate timescales ( $\sim 24\%$  of the sample).

As detailed in Sect. 4.2.2 and Appendix B, a combination of factors such as the sparse sampling of GDR3 and the SF discretisation process makes the SF method inefficient for detecting aperiodic timescales in the range  $\sim 0.3$ –28 d. Hence, if

the GLS method was inefficient at detecting long timescales, this bi-modality could be explained by the detection limits introduced by the combined use of these two methods. Conversely, Fig. 11 shows that the distribution of timescales measured with the GLS method is itself bi-modal, with 28% of GLS-derived values at intermediate-timescales. Additionally, albeit much reduced in numbers, the SF timescale distribution accounts for  $\sim 6\%$  of rotation timescales. Another possible explanation would be that the intermediate-timescale population is the result of the spurious variability discussed in Sect. 3.3. However, we have addressed and mitigated the occurrence of residual spurious variability in our sample in Appendix A.4 and Sect. 4.3, where we discarded sources with timescales in the ranges typically

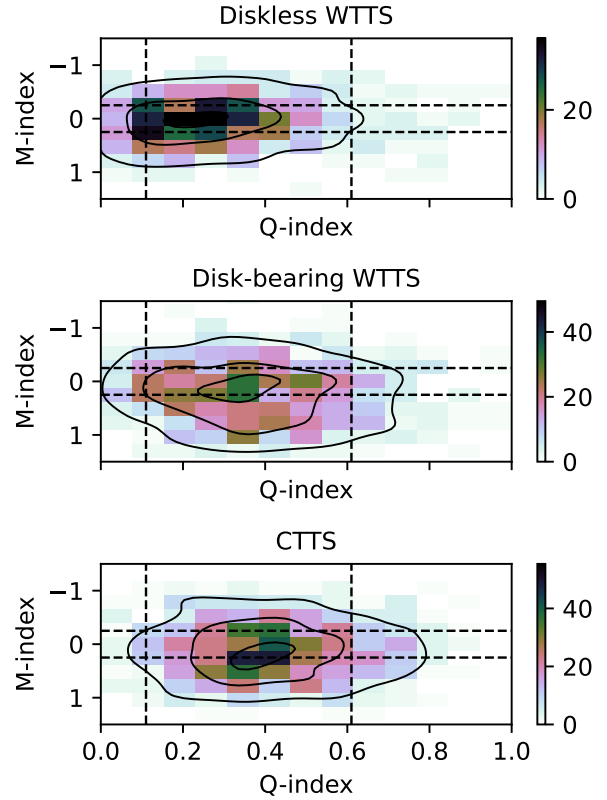


**Fig. 10.** Histograms for the distributions of the Q (top) and M (bottom) indices. L variables are included in the M distribution, but excluded from the Q distribution. The full sample is shown in black, sources with periodic GLS timescales in red, and those with aperiodic SF timescales in blue. The primary and secondary samples appear as magenta dashed and green dash-dotted lines, respectively.



**Fig. 11.** Top: timescale distribution for the full sample (black), 19 345 sources with GLS timescale (red), and 3868 sources with an SF timescale (blue). Bottom: amplitude vs. timescales for the full sample of 23 213 sources with a derived timescale. L variables are excluded.

susceptible to spurious variability. Finally, when comparing our results with previous studies using ZTF light curves (Sect. 6.3.2), we found that those previous studies also find a hint to similar bi-modality in timescales, with  $\sim 74\%$  vs  $\sim 9\%$  incidence of rotation and intermediate timescales. We are therefore confident that the observed bi-modality is real. The two populations also show distinctive behaviour in ptp amplitudes, with rotation timescales showing average amplitudes of 0.18 mag (standard deviation 0.26) and intermediate timescales at 0.29 mag (0.38), and these differences are also supported by a KS test.



**Fig. 12.** Q–M distributions for subsamples of sources with no observational indication of ongoing accretion (WTTs) and absence of a disc as identified from their IR SED (top), non-accreting sources (WTTs) whose IR SED indicates a disc (middle), and sources identified as actively creating (CTTS, bottom).

### 5.3. Asymmetry and periodicity metrics in the context of the observed YSO properties

In this section, we examine our results in the context of quantities related to the physical characteristics of YSOs.

#### 5.3.1. Accreting versus non-accreting sources

In Appendix C.1, we classified sources as actively accreting Classical T Tauri Stars (CTTS) or non-accreting weak-lined TTS (WTTs) based on their  $H\alpha$  emission. In total, we could assign CTTS/WTTs labels to 13 802 GDR3 YSOs, with 7738 in the variable sample. Among 2829 sources in the primary sample with both Q and M derived, 1107 were CTTSs and 1722 were WTTs. Their Q–M distribution is shown in Fig. 12, where we further distinguish between disc-bearing and discless WTTs based on their  $\alpha_{\text{IR}}$  index (Sect. 5.3.2). Sources with  $H\alpha$  information have a lower incidence of aperiodic variability ( $\sim 11\%$ ) compared to the full variable sample ( $\sim 20\%$ ).

The Q–M distribution for discless WTTs peaks at  $Q \sim 0.29$  and  $M \sim 0.04$ , with Q values shifted toward the intersection between strictly and quasi-periodic sources, and a narrower distribution of M values concentrated toward symmetric M indices. The subsample of actively accreting sources (CTTS) is mainly quasi-periodic, with their distribution peaking at  $Q \sim 0.42$  and  $M \sim 0.13$ . Compared to the discless WTTs distribution, CTTSs are shifted towards larger Q values, with a much wider distribution of M values including more significant fractions of dippers and bursters. For disc-bearing WTTs,  $Q \sim 0.38$  and  $M \sim 0.24$ , with the distribution presenting intermediary properties between

the former two groups. Disc-bearing WTTSs show a smaller relative fraction of P sources compared to discless WTTSs as well as an excess of QPDs with higher M values than CTTSs.

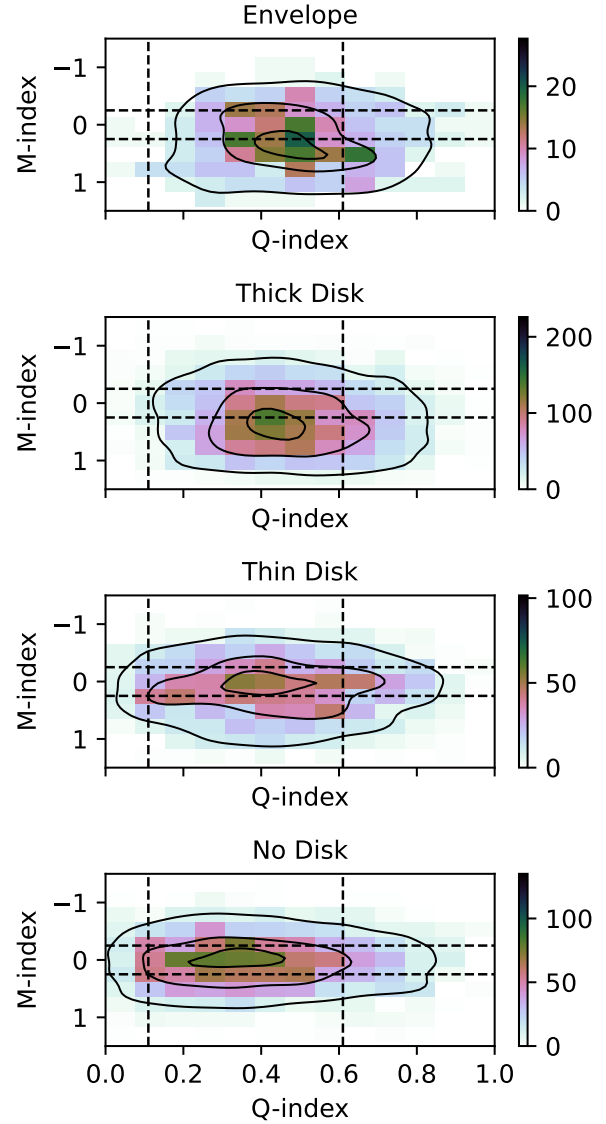
To further support the discussion of the results, Table 1 includes the CTTS frequency defined as the percentage of CTTS sources among all sources with  $H\alpha$  measurements, and interpreted here as a measure of the fraction of sources actively accreting within a sample. We note that  $H\alpha$  measurements are limited to optically visible sources. Consequently, this proxy will inevitably miss deeply embedded sources. Furthermore, we also note that the accretion process and the resulting  $H\alpha$ -emission are also highly variable (e.g. [Sousa et al. 2016](#)). Hence, although strong  $H\alpha$  emission indicates ongoing accretion, the opposite is not always true, as sources may still be sporadically accreting over longer timescales although not in the specific season observed. CTTS fraction must thus be interpreted as a bottom limit of the true fraction of actively accreting sources.

### 5.3.2. Circumstellar discs

To gauge the influence of circumstellar discs on our results, we examined the  $\alpha_{IR}$  indices for our sources. The  $\alpha_{IR}$  index is a measure of the slope of the IR spectral energy distribution (SED) and quantifies the contribution to the SED of infrared radiation re-emitted by material around YSOs. In Appendix C.2, we were able to derive  $\alpha_{IR}$  for 27 653 sources (98% of our final sample). To facilitate our discussion, we used  $\alpha_{IR}$  to classify sources according to the dominant contributor to their SED: 712 sources in the primary sample and 697 in the secondary had a significant envelope contribution; 4296 (primary) and 2131 (secondary) sources had a thick disc; 2021 and 6722 sources had a thin disc; and 2795 and 8836 sources were likely discless. Fig. 13 shows the Q–M distributions for each of these samples. Table 1 includes the frequency of sources with  $\alpha_{IR}$ -based evidence for the existence of an inner disc, which we define as the total number of sources classified as envelope or thick disc, divided by the total number of sources with  $\alpha_{IR}$  index.

KS tests between envelope and thick disc samples suggest that these samples may share the same mother Q–M distribution. However, when examined in terms of the ratio between rotation and intermediate timescales, we find a 0.7 ratio for the envelope sample compared to a 1.0 ratio for the thick disc sample. Generally, Fig. 13 suggests a gradual transition towards a distribution of M indices that favours symmetry and a shift toward enhanced periodicity for more evolved YSO populations.

Several of the variability mechanisms in YSOs stem from the star-disc interaction process, where the young star magnetically interacts with its inner disc. Hence a relevant quantity included in Table 1 is the percentage of sources with  $\alpha_{IR}$  index indicative of the presence of an inner disc, which we define as  $\alpha_{IR} \geq -1.6$  and includes both sources with a thick disc or an envelope. We also report the fraction of sources with evidence for circumstellar material (namely general disc-fraction), which we define as the fraction of sources  $\alpha_{IR} \geq -2.5$  or that had EW( $H\alpha$ ) indicative of accretion (Sect. 5.3.1.). Our  $\alpha_{IR}$  are derived based on heterogeneous data available from previous publications and thus may be biased toward a lack of mid- and far-infrared data, which are only available all-sky within the context of *WISE* ([Wright et al. 2010](#)), with  $\sim 31\%$  of our compiled SEDs including only data below  $5 \mu\text{m}$ . We therefore stress that reported disc frequencies represent a lower bound on the actual frequencies. For example, the thick disc fraction for the full sample goes from 25.5% to 37.3% when considering only sources with SEDs populated beyond  $\lambda \geq 5 \mu\text{m}$ . For a discussion of how using SEDs without

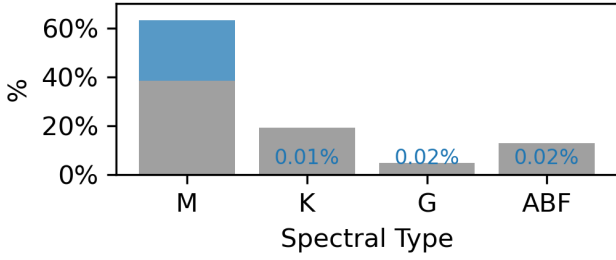


**Fig. 13.** Q–M distributions for subsamples of sources with  $\alpha_{IR}$  index (see Sect. 5.3.2) indicative of contribution of an envelope to their SED (top), a thick-disc (middle-top), a thin disc (middle-bottom), or discless sources (bottom). Colour maps reflect the source count per bin.

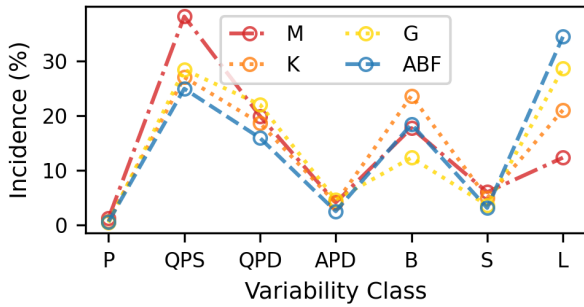
data beyond  $\lambda \geq 5 \mu\text{m}$  affects studies based on  $\alpha_{IR}$ , see [Roquette et al. \(2025, Appendix C\)](#), who report a 25.2% impurity among thin disc sources identified from such undersampled SEDs.

### 5.3.3. Spectral types

Fig. 14 shows the distribution of spectral types for our sample, estimated in Appendix C.3, illustrating that  $\sim 83\%$  of our sample is composed of K and M sources, with a smaller incidence ( $\sim 13\%$ ) of intermediate mass YSOs. In Table 1 we present the incidence of spectral types in each variability class, and in Fig. 15, the incidence of variability classes for each spectral type. The latter, in particular, illustrates two trends as a function of spectral type. Earlier spectral types are more frequently classified as L variables, whereas QPS sources occur more often at later spectral types. We also note that B, QPD and APD also vary with spectral type, although variations are much smaller and show a less clear trend. In our discussion of the results, we group A, B and F type objects as intermediate-mass or early-type



**Fig. 14.** Percentage of sources in the YSO variable sample as a function of their spectral types. Spectral types derived in Appendix C.3 from their GSP-Phot effective temperatures are shown in grey, and in blue if derived from de-reddened  $G_{BP} - G_{RP}$  colours. Bars too small for visualisation have their fractional contribution written in the plot.



**Fig. 15.** Incidence of variability classes as a function of spectral type.

YSOs. G, K, and M type objects are treated as late-type sources, with K and M type sources distinguished when relevant.

## 6. Discussion

The effectiveness of the Q–M metrics for characterising YSO variability has already been demonstrated in the literature with successful applications to both high-cadence space-based and longer-term ground-based observations. In this study, we extend prior work by showing that Q–M metrics can be applied to Gaia’s sparse yet highly precise long-term data. Having established this, we next consider how to interpret our findings within the Q–M morphological classification scheme.

### 6.1. Incidence of morphological classes across surveys

Table 2 compares the population-level frequencies of variability morphologies reported in different works. Morphological classes from the literature are taken as reported, without review of their Q–M thresholds. Some previous studies included extra classes for multi-periodic and non-variables, which we did not consider here and we estimated their class incidences excluding these two categories. We thus stress that these frequencies reflect the incidence of morphological classes among samples of variable stars. Fig. 16 presents amplitudes versus timescales distributions for each morphological class. This includes our primary and secondary samples, as well as the main samples from the literature combined by type of survey, including high-cadence surveys (CoRoT and K2; Cody et al. 2014; Cody & Hillenbrand 2018; Venuti et al. 2021; Cody et al. 2022) and ground-based results (ZTF; Hillenbrand et al. 2022; Wang et al. 2023; Jiang & Hillenbrand 2024). Detailed discussion of each morphological class are later presented in Sect. 6.2.

**Table 2.** Incidence (in per cent) of the variability classes for GDR3 light curves in comparison with previous studies.

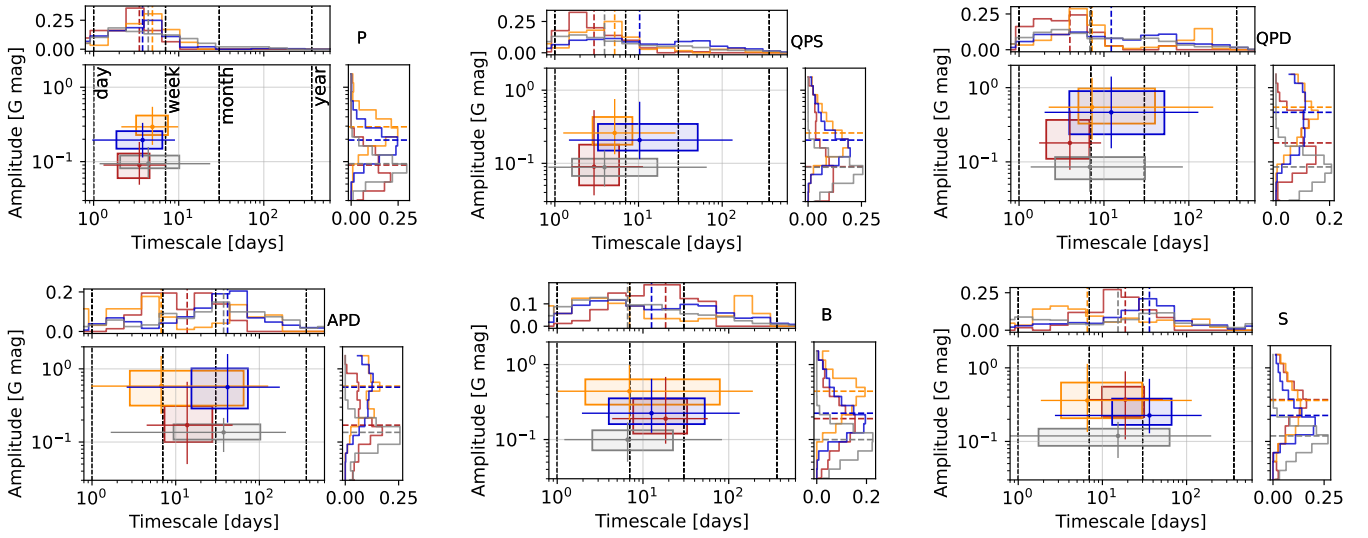
Class	This work			Previous studies <sup>a</sup>	
	Full Sample	Primary	Secondary	High cadence	ZTF
P	1.0	2.0	0.5	6.7	14.5
QPS	33.7	24.0	39.0	28.3	35.6
QPD	19.3	29.4	13.9	16.1	13.8
APD	3.9	8.4	1.5	15.7	10.0
B	18.6	15.3	20.4	17.6	8.9
S	5.3	6.1	4.8	12.8	14.0
EB	0.5	1.0	0.2	0.2	–
L	17.6	13.8	19.7	2.6	3.2 <sup>b</sup>

**Notes.** <sup>(a)</sup>The reported incidences from previous studies refer to the primary variability classifications and were recalculated to account for the total number of variables in P, QPS, QPD, QPD, B, S, EB, and L. <sup>(b)</sup>The L incidence reported for ZTF (3.2%) does not account for sources classified as such in their secondary classifications (~10%).

Discussion of results must acknowledge survey-to-survey differences. First, in terms of sample size: our study included 9824 YSOs in the primary and 18 355 in the secondary sample, while ZTF and high-cadence surveys included 1031 and 549 sources, respectively. With much smaller sample sizes, previous studies relied heavily on visual inspection of light curves for classification, with Q–M thresholds adjusted by eye to maximise correspondence with visually assigned morphological classes. Although infeasible in large-scale surveys such as Gaia, this empirical approach adopted in previous studies suggests that survey-to-survey adjustments of Q–M thresholds can help account for surveys’ sampling limitations and may improve purity within morphological classes. In Sect. 4.4, we adopted Cody et al. (2014) original Q–M thresholds for classification. As further discussed in Sect. 6.2, our results suggest that these thresholds led to some contamination of symmetric light curves into asymmetric classes within our sample. With the ground truth of visual inspection further hindered by the sparse light curves of Gaia, an alternative approach would be to review Q–M thresholds based on simulated light curves. Although beyond the scope of the present study, our team has been exploring this approach as part of our preparations for Gaia DR4.

High-cadence studies typically have photometric precision of ~0.001 mag, compared to ~0.2 mag for ZTF, implying that the ZTF sample is biased towards higher-amplitude variables. In contrast, our GDR3 sample has typical photometric uncertainties  $\approx 0.001$ – $0.005$  mag for  $G < 17$  mag, and  $\lesssim 0.02$  mag up to  $G \sim 19.5$  mag. Based solely on photometric precision, our primary sample should most closely match ZTF, while for  $G \lesssim 17$  mag, the full sample should best resemble high-cadence surveys. The incidence of morphological classes for the former is shown in Table 2, the latter is omitted, as the results were roughly equivalent to the full sample. However, as discussed in this section, we speculate that cadence and sample selection play a stronger role in explaining different results at the population level across surveys.

There are also source selection differences, where high-cadence surveys examined only disc-bearing sources, while ours and ZTF results include all YSO types. Moreover, previous studies focused on individual star-forming regions, with much narrower age distributions in comparison with the GDR3 all-sky sample. Notably, Cody et al. (2022) attempted to interpret



**Fig. 16.** Box plots of the timescale–amplitude distributions are shown for six morphology classes defined by the Q–M indices, along with normalised histograms for each parameter. The primary sample is plotted in blue and the secondary sample in grey. High-cadence space-based sources (CoRoT, K2) and ground-based ZTF results are shown in red and orange.

the frequency distribution of morphological classes in different star-forming regions in light of their age. In our case, this type of interpretation is hindered by the lack of previous constraints in the age distribution of the GDR3 YSO catalogue. However, our investigations of  $\alpha_{\text{IR}}$ -indices in Sect. 5.3.2 revealed that 42% and  $\sim 32\%$  of our sample of variables had no indication of the presence of a disc or only a thin-disc contribution to their SEDs.

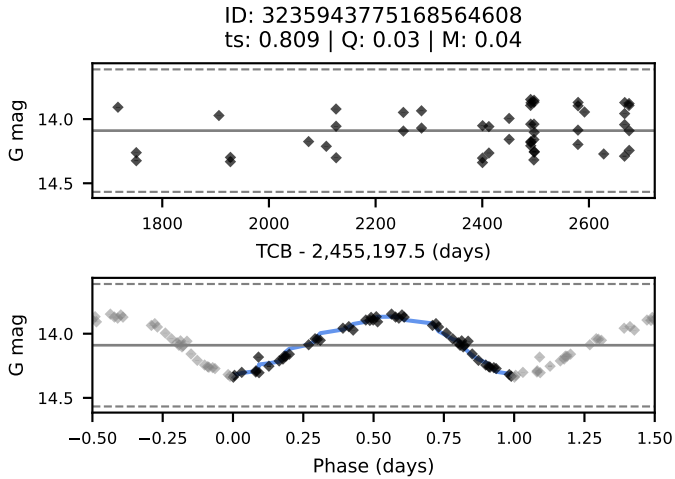
Given the enhanced fraction of discless YSOs in our sample, cool spot variability, typically associated with P and QPS variability, is expected to be the most common in our sample. Although QPS is in fact the most common variability class, Table 2 shows that P frequency is much lower than in previous studies. Beyond cadence and underlying noise, we further discuss in Sect. 6.2.1 that this can be attributed to cool-spot variability being unstable at timescales of the order of GDR3 observations. We note that cool-spot variability is a central feature of “solar-like variables”, which are themselves a subgroup of GDR3 variability classes, with Distefano et al. (2023) reporting that 53–72% of their 474 026 variables show unstable spot-variability (amplitude variation) during the observed time covered by GDR3. This helps explain not only our under-detection of P variables compared to previous studies, but also contributes to our enhanced detection of QPS variables. Table 2 also indicates that ZTF studies, which also include discless sources, show a much larger sample of P and QPS sources in comparison to high-cadence surveys. However, ZTF studies use much larger Q thresholds to account for the higher noise from ground-based observations.

The solar-like variables reported by Distefano et al. (2023) focus on late-type stars, including some relatively young objects, while being disjoint from the GDR3 YSO sample. This suggests that a certain fraction of young P and QPS variables may be lost to their sample. To quantify sources lost to the GDR3 sample of solar-like variables, we further examined the NEMESIS YSO catalogue for the Orion Star Formation Complex (Roquette et al. 2025). Of 27 879 YSO candidates, the NEMESIS sample included 5231 sources in common with the Gaia YSO sample, of which 111 were classified as P and 2131 as QPS, and a further 152 sources in common with Gaia solar-like variables, suggesting that we miss about  $\sim 7\%$  of cool-spot variables. Similarly,

EBs are also a subclass of variables in Gaia, with Mowlavi et al. (2023) reporting more than 2 million EB candidates as part of DR3. A similar comparison to the NEMESIS catalogue suggests that our EB detection rate might be reduced by half.

We also point to survey-to-survey differences in the time span of observations. High-cadence surveys cover between 40 and 80 days of observations, while ZTF surveys cover 800–1500 days, and GDR3 750–1000 days. These differences imply that quasi-periodic phenomena at intermediate timescales may be perceived as aperiodic by studies with reduced time span. Moreover, L variables are defined as sources with timescales larger than 20–40 d in high-cadence studies, 100–250 d in ZTF, and we used 350–500 d in the present study. L variable fractions are thus fully survey-dependent. Even if removed from our class incidence comparison, their impact on the definition of other morphological classes is still worth mentioning. First, L variables could be operationally defined as aperiodic, since their time scale was not resolvable over the duration of a survey. Although their lack of reliable timescale implies that their Q indices cannot be estimated, accounting for their M indices would increase class frequencies by 3.8% for B, 8.6% for S, and 5.3% for APDs – thus bringing our class incidence for S and APD variables to levels similar to those of ZTF studies. Second, the timescale limit for the L variable definition indirectly determines how many folds are used in the waveform estimation required to derive the Q index. With our threshold of half of the total duration of observations, only one fold is included for sources with the longest timescales, which may lead to over-fitting of estimated waveforms, resulting in deflated Q indices, and causing aperiodic phenomena to be miss-detected as quasi-periodic.

Some previous studies attributed multiple morphological classes to multi-mode variables (e.g. Cody & Hillenbrand 2018; Cody et al. 2022; Hillenbrand et al. 2022) also including an additional class, multi-periodic, to account for variables with multiple periods. In Table 2, multi-mode variables were represented by their primary morphological class. However, with our focus on the dominant variability mode, we acknowledge that undetected multi-mode variables in our sample may display inflated Q indices. Beyond multiple variability mechanisms



**Fig. 17.** Example of a P variable. Top: Full  $G$  light curve. Bottom: Phase-folded light curve. The blue line represents the mean waveform used in the Q calculation, the grey line indicates the mean magnitude, and the dashed lines mark the  $3\sigma$  variation bounds. Same descriptions applies to all following figures until Fig. 24.

occurring simultaneously, the longer time span of surveys like Gaia and ZTF may also facilitate the observation of sources with alternating dominant variability phenomena. For example, this can result in symmetric M indices derived from light curves that shift from bursting to dipping behaviour (e.g., Fig. 22).

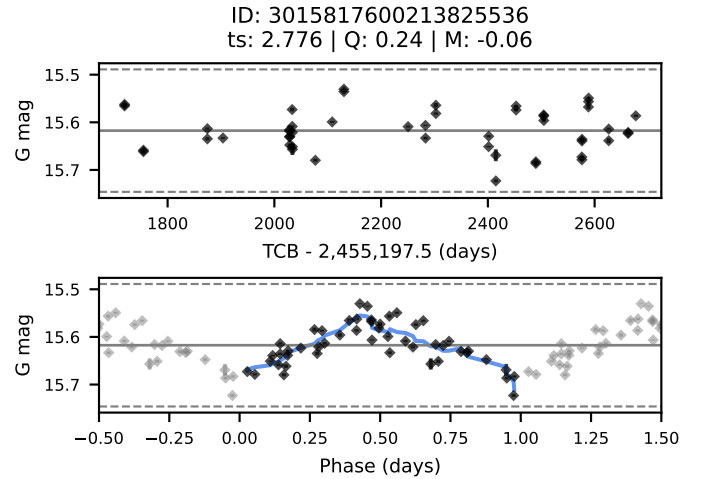
Our total fraction of dipping variables (QPD and APD) matches that reported in ZTF studies, but is lower than in high-cadence surveys. Nevertheless, as already noted, our APD fractions are underestimated due to our definition of L variables. Furthermore, in Sects. 6.2.3 and 6.2.4 we discuss evidence that our sample of dippers includes a different underlying population of YSOs in comparison to the high-cadence surveys. Similarly, B variables also include different variability mechanisms, and in Sect. 6.2.5 we discuss evidence that the Gaia sampling yields the detection of a distinct population of bursting variables in comparison with other surveys.

## 6.2. General properties of the morphological classes

In this section, each variability morphological class is discussed within the context of its expected physical origins and its typical timescale and amplitude.

### 6.2.1. Strictly periodic

Strictly periodic variables (P, Fig. 17) are usually associated with rotational modulation by magnetically induced cool spots at the stellar surface in sources viewed close to edge-on ( $\sim 40\text{--}90^\circ$ ; Cody et al. 2022). Fig. 16 shows that the timescale distribution for P sources is consistent across studies and dominated by rotational timescales. The variability amplitude in our primary sample (blue) is higher than that reported in high-cadence surveys (red) but slightly lower than in ZTF studies (orange); the secondary sample (grey) exhibits amplitude levels comparable to those of high-cadence studies. Differences in amplitude distributions can be attributed to selection effects in different surveys. Typically, cool spots have lifetimes ranging from 1 to 50 rotation cycles (Basri et al. 2022). Higher-cadence surveys, which usually span 40 to 80 d, provide excellent phase coverage before significant changes in spot coverage occur. In Gaia, sparse



**Fig. 18.** Example of a QPS variable.

sampling and long-term coverage implies that hundreds of rotation cycles take place before achieving sufficient phase coverage, thus biasing the sample of P sources towards sources with larger and more stable spot coverage (and large amplitudes). In contrast, the higher noise level of ground-based observations allows only the higher-amplitude variables to be detected. As a result, Fig. 16 shows that the ZTF distributions of all classes are shifted toward higher amplitudes compared to other surveys.

The properties of our P variables support the hypothesis that their variability is primarily driven by rotational modulation linked to relatively stable cool spots. Their average amplitude is among the lowest of all morphological classes, and the ratio between rotation and intermediate timescales (Table 1) shows that this sample is dominated by rotational modulation. P variables are 82.95% composed of YSOs without evidence of a thick disc or envelope, and 90.2% composed of K- and M-type stars. Furthermore, 86.27% of P sources with  $H\alpha$  information are WTTS.

### 6.2.2. Quasi-periodic symmetric

Multiple physical origins have been proposed to explain QPS variables (Fig. 18), which are the most common in our entire sample. Although generally linked to rotational modulation of stellar spots, high-amplitude QPS have been tentatively associated with variability caused by large hot spots induced by accretion, in combination with underlying lower-amplitude accretion variation (Venuti et al. 2015; Cody & Hillenbrand 2018), while low-amplitude QPS are thought to originate from changes of cool spot coverage due to magnetic activity (Cody et al. 2022). Considering that cool spots are the only variability driver expected in discless sources and that disc-bearing and actively accreting sources are likely to exhibit accretion-induced hot spots, this view is generally supported by our sample. We find that the average distribution of amplitudes is lower for discless and thin-disc QPS (0.11 mag) compared to sources with envelope or thick disc (0.30). Similar differences are found when comparing the amplitudes of the WTTS QPS (0.11) with the CTTS QPS (0.36). The same trend is verified by comparing our primary and secondary samples, where the CTTS frequency varies from 36.6% to 6.1%. The link of QPS with rotational modulation by spots is further reinforced by the observation that M-type QPS have an excess of short timescales, with 60% of them displaying timescales of  $\lesssim 5$  days compared to  $\sim 30\%$  for other

spectral types. This is in accordance with the expectation that M-type stars tend to be faster rotators than their more massive counterparts (e.g. [Cieza & Baliber 2007](#)).

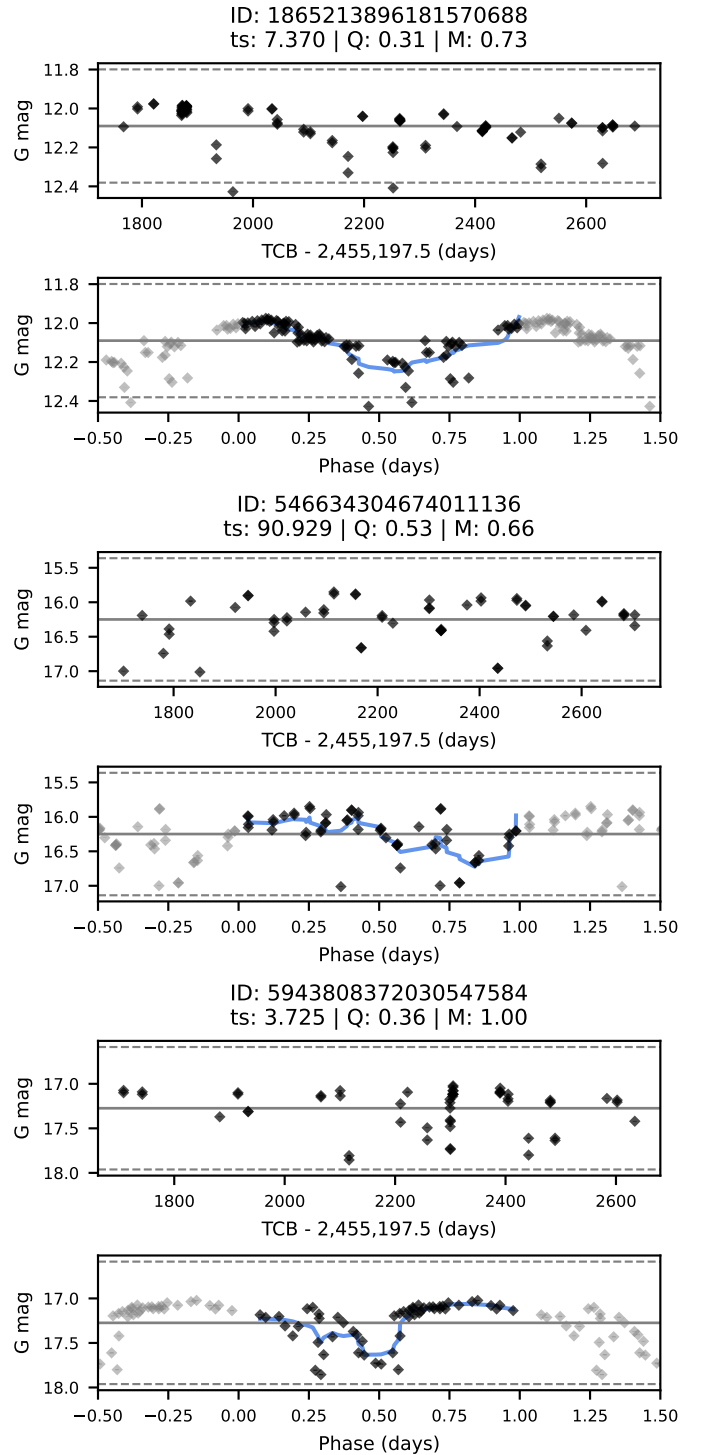
Our QPS-variables include a population of sources with intermediate timescales, which has not been found in previous studies. Fig. 16 shows that for the primary sample, the timescale distribution is bimodal, with 1.3 times as many rotation timescales as intermediate ones. A comparison of their Q index distributions reveals significant differences (with KS test support), where the rotational-timescale QPS are shifted toward higher periodicity with an average Q of 0.34 compared to 0.44 for intermediate-timescale QPS.

Visual inspection of intermediate-timescale QPS light curves revealed that many of these sources have limited phase coverage when folded to their characteristic timescale. This limitation is inherent in surveys with sparse sampling as Gaia, where some phase-folded light curves include large gaps with no data. These gaps, in turn, bring us to the limits of the method for waveform estimation, causing waveforms to be over-fit to only a fraction of the phase, artificially shifting Q indices towards lower values and causing misdetection of aperiodic sources as quasi-periodic. With the Gaia scanning law, the timescales that yield gaps in phase coverage vary with sky position, making modelling and masking these timescales from our analysis impractical. To mitigate this issue, future studies should consider the complementation of Q–M analysis with an additional variability metric sensitive to the degree of phase coverage as part of the procedure for the derivation of the Q index (e.g. [Saunders et al. 2006](#)). We also considered adjusting the quasi-periodic/aperiodic Q index threshold to mitigate this effect, but verified that a lower Q index limit would result in reliable quasi-periodic light curves with rotation timescales being misclassified as aperiodic.

### 6.2.3. Quasi-periodic dippers

The phenomenology AA Tauri variables ([Bouvier et al. 1999](#)) is often linked to QPD light curves (see Fig. 19), where a misalignment between the rotation and magnetic field axes yields the formation of wraps, with disc material lifted above the disc mid-plane and close to the co-rotation radius, introducing quasi-periodic occultations at rotational timescales in systems observed close to edge-on ([Romanova et al. 2013](#); [McGinnis et al. 2015](#)). Among the primary sample, 62.09% of QPDs have an inner disc, and 40.60% have evidence of ongoing accretion (CTTS classification), supporting the association of this class with physical phenomenology related to the presence of a disc and ongoing accretion. However, a direct association of these with the AA Tau prototype is hampered by two factors.

Firstly, a fraction of our low-amplitude QPDs may be misclassified QPS. About half of our QPDs are in the secondary sample, which shows a much lower disc (13.71%) and CTTS (4.47%) frequencies, with both Q and M distributions shifted towards more symmetric M and more periodic Q indices compared to the primary sample. Fig. 16 shows that these have systematically lower amplitudes than QPDs from high-cadence surveys. The M index distribution for the secondary QPD sample peaks near the threshold for dipper selection, with 90% of the M indices below the median value for the primary QPD sample. Additionally, although statistically different in shape, the distributions of amplitudes and Q index for the secondary QPD and secondary QPS samples have similar averages and standard deviations, as well as similar disc and CTTS frequency. Finally, visual examination of the light curves also suggested that a higher M index threshold might be necessary to enhance the



**Fig. 19.** Examples of QPD variables at rotational timescales (top and bottom) and intermediate timescale (middle).

purity of the dipper variable samples, where a reviewed threshold would identify a large fraction of secondary QPD as QPS.

Secondly, because the observation time span of Gaia is much longer, QPDs may be sensitive to other dimming variables beyond AA Tau. Fig. 16 reveals a bimodal distribution for QPDs, with  $\sim 35\%$  of QPDs showing intermediate timescales. This contradicts the predictions of the AA Tau phenomenology, where the occulting material is anticipated to be near the co-rotation radius. The primary sample of QPDs with intermediate

timescales shows high disc and CTTS frequencies (65.09, and 47.35%), suggesting a physical origin also associated with the presence of a disc and ongoing accretion. Statistical tests for distribution shape, mean and variance differences for the distributions of Q indices and amplitude support the hypothesis that our primary QPD sample is composed of two distinct populations of variables with rotation and intermediate timescale, where the rotation timescale QPDs have lower amplitude on average (0.29 mag) and are more periodic (lower Q), while the intermediate timescale QPDs tend to have larger amplitude on average (0.44 mag) and reduced periodicity in comparison. We note that this second population of intermediate-timescale QPDs is also present in the ZTF sample, which has a time span close to that of Gaia, suggesting that this second type of QPDs is absent from high-cadence surveys due to their limited total time span.

While QPDs at rotational timescales can be associated with AA Tau type variability, those with intermediate timescales could be related to occultation by disc material located farther away than the inner disc. One possible prototype for this variability behaviour could be the young star RW Aur (Cabrit et al. 2006; Facchini et al. 2016). A proposed phenomenological explanation for the fading of RW Aur would be a tidally truncated disc in which a flyby removed the outer part of the disc, leaving a tidally disrupted ‘arm’ that induces occultations of stellar light when observed from moderate viewing angles (e.g. Rodriguez et al. 2013). We note that although RW Aur is best known for its 2 mag deep fading events in 2010 and 2014, lower amplitude variability could be expected under the same phenomenology (e.g. Fischer et al. 2023). A different possible phenomenon could be disc instabilities as a trigger of dense disc winds responsible for enhancing line-of-sight extinction above the disc mid-plane, which, due to their radial location, would cause orbital modulation at longer timescales (e.g. Bouvier et al. 2013). Further constraints on the colour behaviour during dipping features could help further probe QPDs.

#### 6.2.4. Aperiodic dippers

With the highest inner disc and CTTS frequencies (56.94 and 39.64%), APD variability (Fig. 20) is likely associated with the star-disc interaction process. Previous studies have also linked APDs to the physical phenomenology of AA Tau. However, while QPDs require star-disc systems that undergo stable accretion to allow dust to be lifted above the disc midplane long enough for the occultations to be observed as quasi-periodic, in APDs, an unstable accretion regime will instead result in stochastically distributed aperiodic dippers through the light curve (Kulkarni & Romanova 2008; McGinnis et al. 2015). Nevertheless, material occulting the star is still expected to be at around the co-rotation radius, thus aperiodic occultation features are still expected to have durations around rotational timescales. In contrast, Tables 1 and Fig. 16 reveal that APDs show intermediate timescales 2.26 times more often than rotational ones.

We note that a meaningful interpretation of aperiodic timescales can be tricky. Fig. 16 shows that APDs from high-cadence surveys have timescales between a week and a month, with these surveys being limited to a time span of 40–80 days. If their APDs are in reality an undersampled quasi-periodic phenomenon occurring over timescales longer than their time span, these will be detected as quasi-periodic by surveys like Gaia or ZTF, which explains the population of intermediate timescales among our QPDs, and the reduced population of rotation timescales in our primary APDs sample. Second, most aperiodic light curves had their timescale determined by the SF

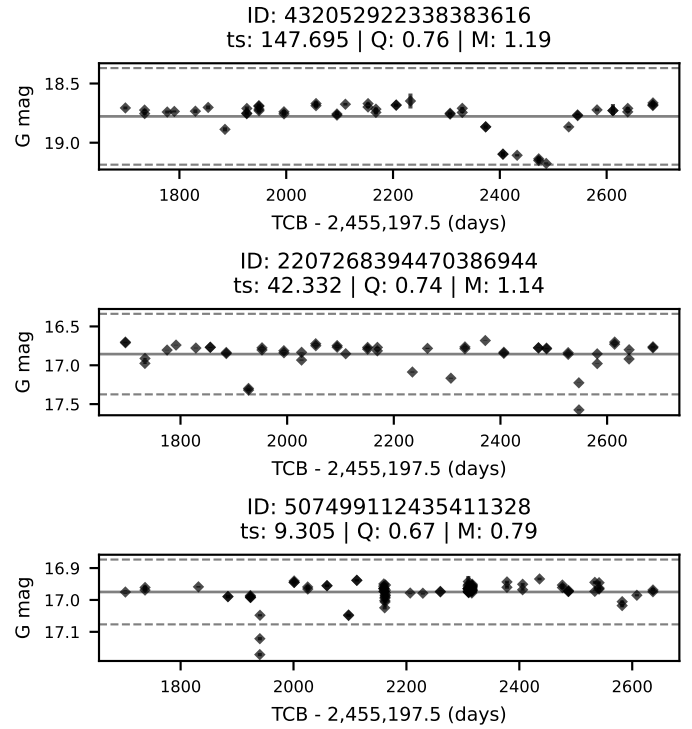


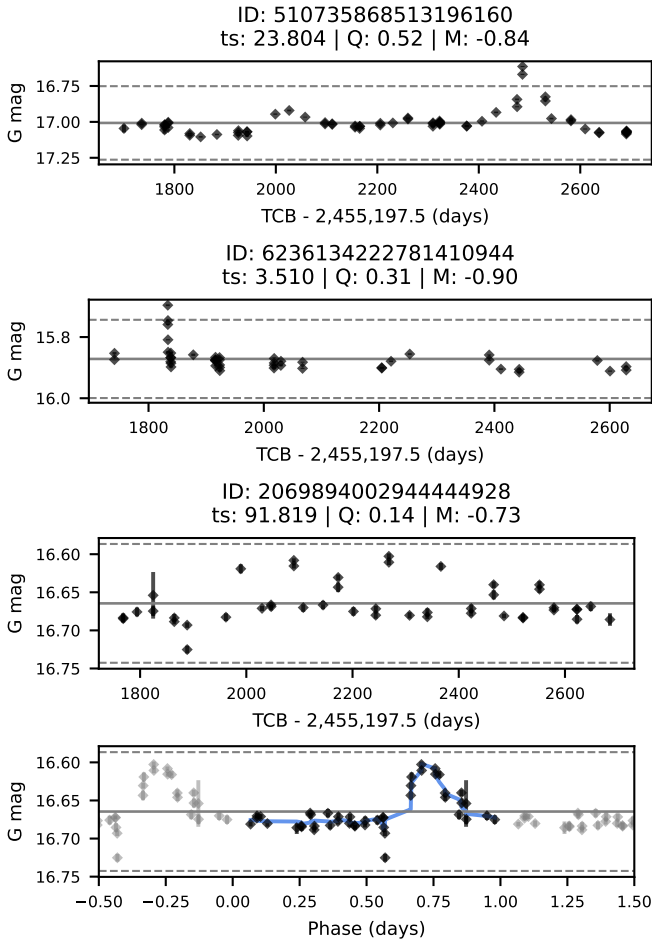
Fig. 20. Examples of APD variables.

in Sect. 4.2.2. SF aperiodic timescales measure the time it takes for the dominant variability feature in the light curve to reach its maximum amplitude, thus reflecting the duration of the variability feature, rather than the frequency of occurrence of variability over the span of observations. For example, the bottom panel of Fig. 20 shows an APD light curve potentially linked to the AA Tau phenomena, where the 9.3 d derived aperiodic timescale reflects the time to reach the maximum amplitude (0.24 mag) in the main dip visible in the light curve. In the top panel of the same figure, we show a second example where the timescale of 147.7 days reflects half of the duration of a main fading event with amplitude 0.53 mag. Combined with the large amplitudes found among our primary sample of APDs, the large disc and CTTS frequencies and large aperiodic timescales support the suggestions that beyond AA Tau-like light curves, APDs at intermediate timescales may be a population of moderate-amplitude RW Aur fading light curves (e.g. Petrov et al. 2015).

With 8.3% of APDs having estimated spectral types between AB and F and 87% of these having a disc, a possible explanation for these earlier type APDs could be the phenomenology behind UX Ori variables (Natta et al. 1997). These YSOs have their variability attributed to large-scale disc perturbations causing a puffing-up of disc material close to the dust sublimation radius (Shulman & Grinin 2022), which, when observed almost edge-on, results in sporadic and large amplitude (1–2 mag) fades. Nevertheless, we verified that although the distributions of amplitudes of earlier and later type APDs are statistically different, the former have much lower typical amplitudes ( $0.36 \pm 0.04$  mag) compared to the latter ( $0.54 \pm 0.02$  mag).

#### 6.2.5. Bursters

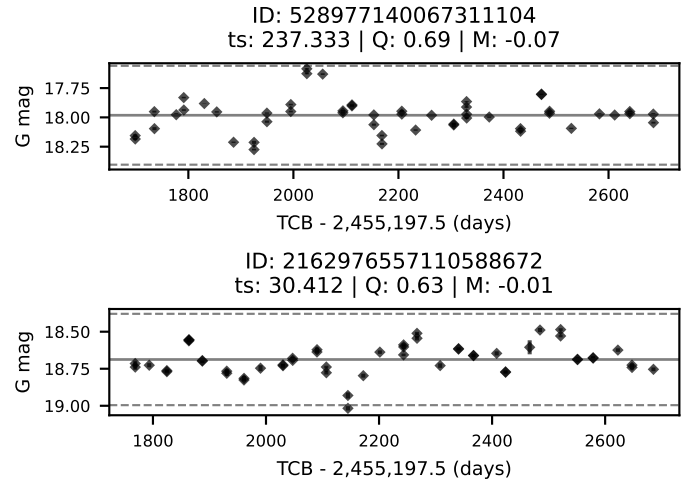
Based on correlations with various accretion proxies, from UV excess to variability in colour behaviour and  $H\alpha$ , B variables were previously associated with high levels of accretion



**Fig. 21.** Examples of B variables with an intermediate duration burst (top), a short duration burst coinciding with a cusp (middle) and a quasi-periodic bursting behaviour (bottom, also shown phase-folded).

(Stauffer et al. 2014; Cody et al. 2017; Venuti et al. 2021; Hillenbrand et al. 2022). In our sample, this association with star-disc-interaction phenomenology is only supported for sources in our primary sample, for which the inner disc and CTTS frequency were 45.18% and 45.49%, while for the secondary sample these are 12.22 and 7.04% respectively. Among these with evidence of a disc or ongoing accretion, we note that the earlier and later type sources share statistically different distributions of amplitude, with the distribution of earlier types (mean:  $0.57 \pm 0.06$  mag) shifted towards higher amplitudes in comparison with later type sources ( $0.38 \pm 0.02$  mag).

As defined by Cody et al. (2014), the B class might constitute a family of bursting behaviours, rather than relating to a single physical mechanism. In a study focused solely on B variables, Cody et al. (2017) finds varied bursting behaviour, with some sources showing a continuous series of bursts, while others presenting short-lived burst episodes on top of underlying irregular variability (see also Venuti et al. 2021). This family of bursting behaviours is verified among our primary sample, and illustrated in Fig. 21, where we show examples of isolated bursts, including both short-lived (GDR3 510735868513196160) and longer-lived bursts (GDR3 6236134222781410944). Fig. 21 also includes an example of a quasi-periodic B variable with intermediate timescale, GDR3 2069894002944444928, which may be related to the phenomenology behind the DQ Tau system (Tofflemire et al. 2017), where two stars orbit each other in a

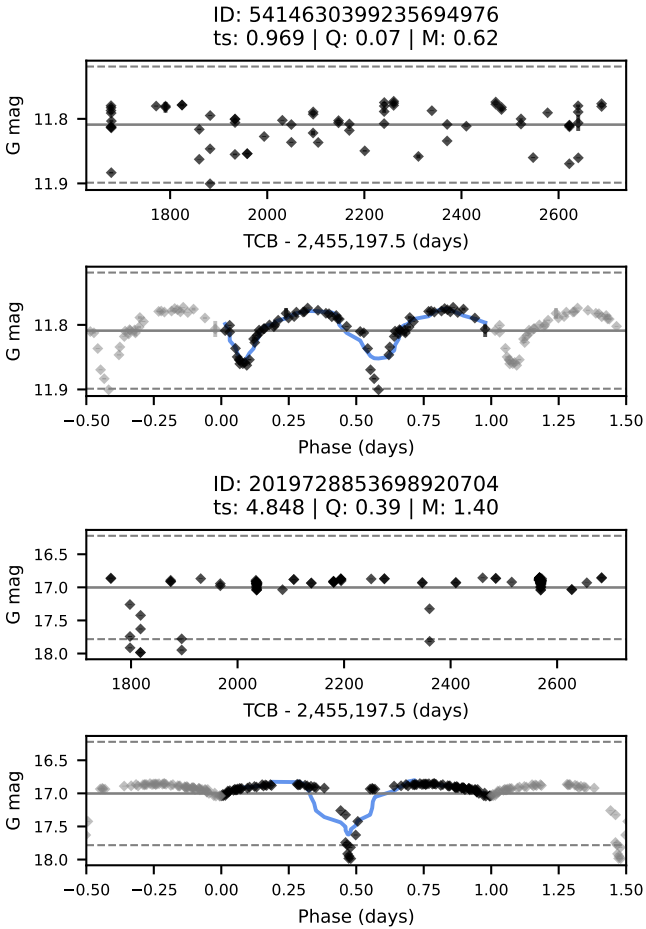


**Fig. 22.** Examples of S variables without preference for bursting or dipping (top) and alternating between dipping and bursting (bottom).

highly eccentric orbit and their magnetic interaction with circumbinary and circumstellar material yields enhanced accretion near periastron, resulting in the observation of regular bursts. Following the discussion on outlier removal in Sect. 3.4 and Appendix A.4, we note that the sparse cadence of Gaia is expected to yield an under-detection of variables with short-lived bursts. As reported by Cody et al. (2017), only 1 burst out of 29 monitored with high-cadence K2 light curves showed burst durations longer than 10 days. If observed by Gaia, such a sample would likely have burst events ruled out as outliers, unless the bursting feature coincides with a cusp, as in the case of a short-lived burst variable GDR3 6236134222781410944 shown in Fig. 21. Finally, we did not find examples of extreme burst events, such as a FUor or EXors. However, we verified the existence of some sharp long-term bursts among our long-trend variables (e.g., Fig. 24).

## 6.2.6. Stochastics

The light curves of S variables (Fig. 22) are dominated by irregular variability without detectable periodicity or a clear preference for bursting or dimming behaviour. Fig. 16 shows that even our primary sample shows systematically lower amplitudes compared to previous studies. Previous high-cadence surveys associated S-variables with a combination of the erratic routine variability introduced by accretion, combined with viewing angles close to face-on, which hindered the rotational modulation of accretion-induced variability (Stauffer et al. 2016). This scenario is supported by our primary S sample, which has inner disc and CTTS frequencies of 43.26 and 49.04%, respectively, along with 2.3 times as many intermediate timescales as rotational ones. However, visual inspection of S light curves shows that a fraction of those are multi-mode variables with alternating bursting and dipping behaviour across the observed time (see Fig. 22 bottom panel). In contrast, our secondary sample has very low inner disc and CTTS frequency (18 and 7.36%), with a wide spread of timescales and no clear preference for rotational or intermediate timescales. However, when disaggregated by amplitude, this sample reveals an excess of sources with low-amplitude (below 0.1 mag) fast rotation timescales (below 5 days), which may suggest a link between this sample and cool spot-dominated light curves close to the Gaia detection limit.



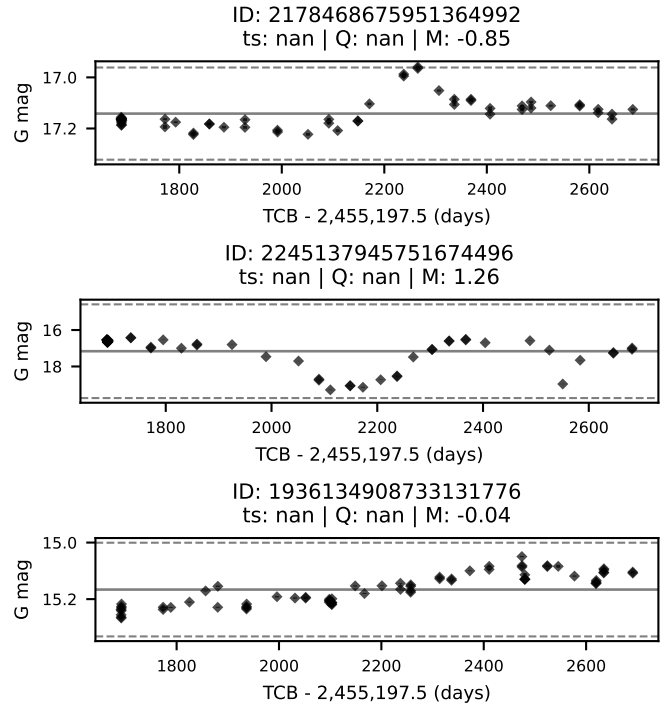
**Fig. 23.** Examples of EB variables, including a QPD misclassification (bottom) due to timescale detection of a harmonic instead of the true period. The phased light curves are shown folded to their true period.

### 6.2.7. Eclipsing binaries

The EB class in [Cody et al. \(2014\)](#) targeted prototypical eclipsing binary light curves with strictly periodic asymmetric light curves. We found 129 EB sources. A search for counterparts in the Simbad Database ([Wenger et al. 2000](#)) reveals that only 4 of our EB sources were previously identified as such in the literature. Although sources classified as EB generally followed the prototype of strictly periodic eclipses, our automated method for timescale derivation using the GLS (Sect. 4.2.1) often detects harmonics of the true period as the best timescale. We show two examples of EB light curves in Fig. 23, where in both cases the GLS returned half of the true period as the best timescale. In the case of the source GDR3 5414630399235694976 (top plots), secondary and primary eclipses are relatively similar in depth, with the waveform estimated based on the biased timescale representing the true waveform relatively well and yielding a very low Q value. However, in cases like the source GDR3 2019728853698920704 (bottom plots), as the secondary eclipse is much shallower, waveform estimation based on the harmonic of the true timescale inflates Q values, resulting in the miss-classification of the source as a QPD.

### 6.2.8. Long trends

With timescales long for a relevant waveform to be derived from the phase-folded curve, L variables lack a Q index, but we still



**Fig. 24.** Examples of L variables for a bursting (top), dipping (middle), and symmetric (bottom) source.

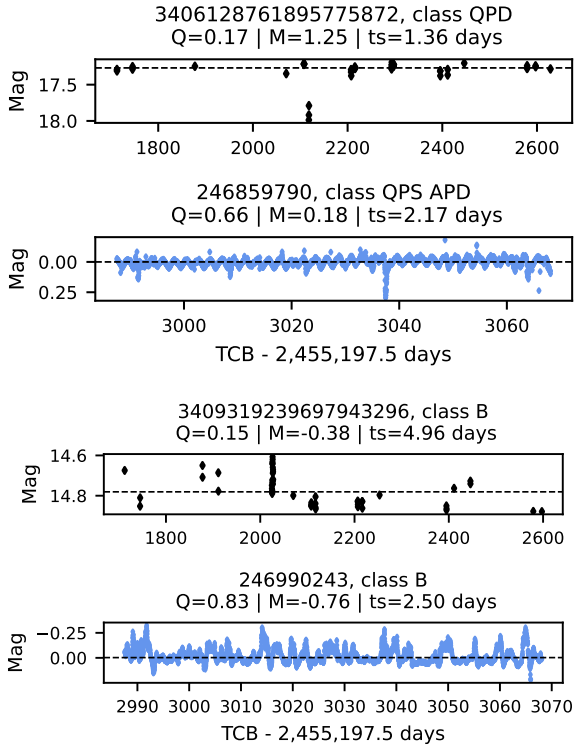
computed their M index. In the primary sample, 21% Ls have bursting behaviour and 30% have dipping behaviour. For the secondary sample, 21% Ls have bursting behaviour, and 26% have dipping behaviour. L variables tend to have larger amplitudes than the other classes, which is expected since there is an empirical correlation between the timescale of an event and its amplitude (See Fig. 3 of [Fischer et al. 2023](#)). Although we do not verify any preference for bursting or dipping as a function of their spectral types, we find that earlier type sources have typically much lower amplitudes (mean:  $0.095 \pm 0.004$  mag) compared with later type sources ( $0.187 \pm 0.005$  mag).

### 6.3. Comparison across surveys

To further investigate the consistency of Q and M indices across different surveys, we cross-matched our variable sample with previous Q–M studies using K2 ([Cody et al. 2022](#)) and ZTF ([Hillenbrand et al. 2022](#); [Wang et al. 2023](#); [Jiang & Hillenbrand 2024](#)). Figs. 25 and 26 show examples of light curves of sources with observations in multiple surveys. Q–M indices from the literature are adopted as reported in the source papers. None of these surveys included seasons of observations in common.

#### 6.3.1. High-cadence space-based surveys

We found only 13 sources in common between our variable sample and the sources observed by [Cody et al. \(2022\)](#) with K2 in Taurus (see Fig. 25). Hence, we limit our comparison to a qualitative discussion. Since K2 and GDR3 have very different cadences and time spans, our comparison focuses on their M index behaviour, where classification was considered equivalent if the source was detected as dipping, symmetric or bursting in both surveys. Although variability amplitudes were generally consistent across surveys, Q–M indices often had large enough differences to produce distinct morphological classes.

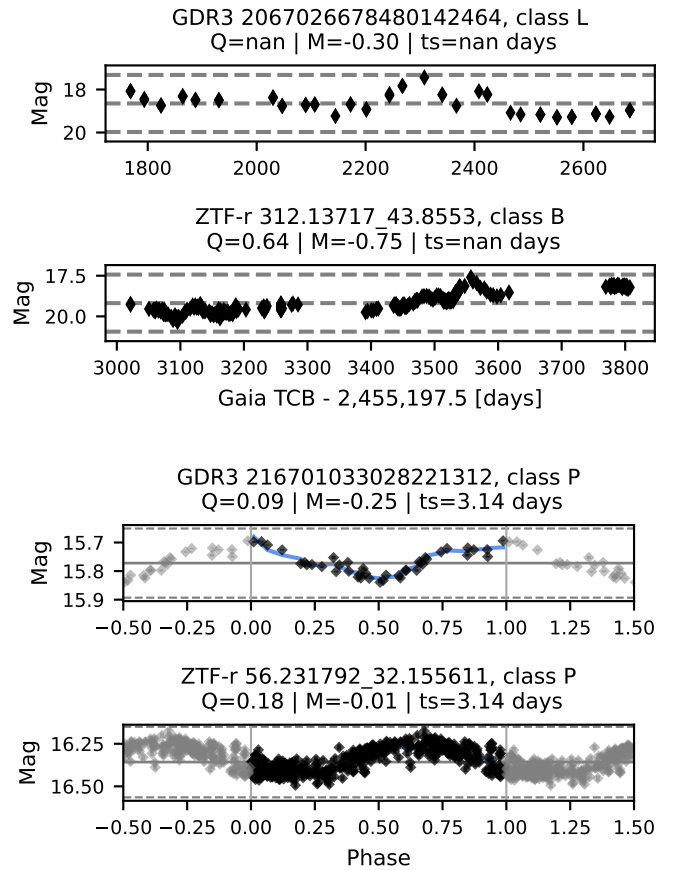


**Fig. 25.** Comparison between GDR3 (top) and K2 (bottom) light curves. The horizontal dashed black lines show the median magnitude, whereas average magnitudes were subtracted from the K2 light curves.

Of the three symmetric sources in K2, only one was also symmetric in GDR3, while the others were dipping. However, since these surveys cover different seasons of observation, it is impossible to infer whether discrepant classifications result from a change in dominant variability mode or the result of different cadences of observations. Three B-variables in K2 were also B in GDR3. One of these sources is GDR3 3409319239697943296 (Fig. 25), which, although bursting in both surveys, was quasi-periodic in our study and aperiodic when observed by K2. Of the four dipping variables in K2, three were also dipping in GDR3. Cody et al. (2022) searched for multiple variability modes by computing  $Q-M$  twice: first on the original light curve and then on the residual data. Two out of three multi-mode sources in common with GDR3 have one of their morphological classes matched to GDR3. An example is shown in Fig. 25, where the QPD source GDR3 3406128761895775872 was attributed to the QPS and APD classes in the K2 analysis.

### 6.3.2. Comparison with ZTF

Light curves observed by ZTF have roughly the same duration as GDR3 but are noisier due to ground-based observations. With 165 sources in common, Fig. 27 compares the  $Q-M$  distribution between surveys, with 25 sources omitted as they were classified as L variables (no  $Q$  index) in the present survey. At the population level, the two surveys yield broadly consistent results, particularly for light curve symmetry with very few cases where variables were detected in radically different behaviours, for example, switching from bursts to dips. For periodic and quasi-periodic sources, we find that differences in derived timescales were either due to harmonics or to one-day aliasing caused by ZTF ground-based observations. For aperiodic sources, timescales differed substantially, but this was explained by the distinct methods used to derive them. Given the

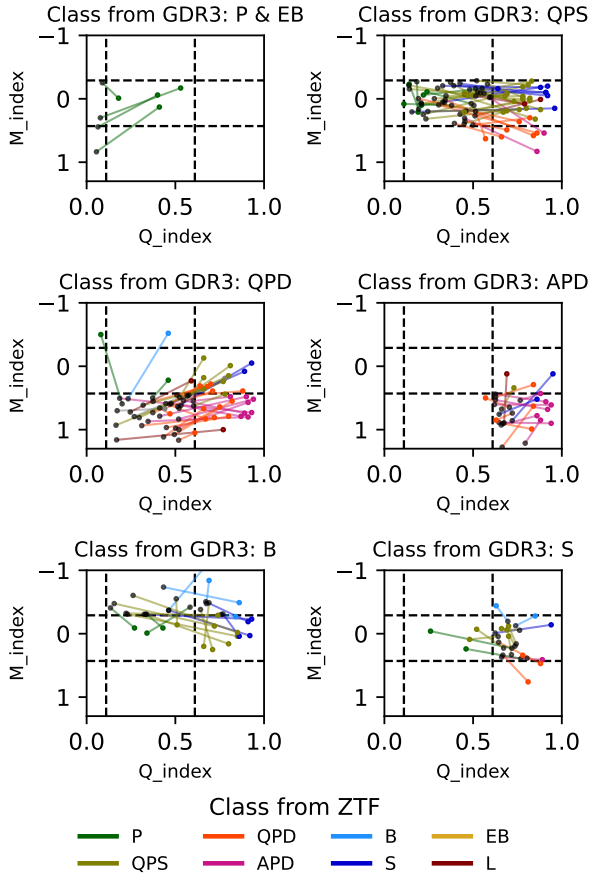


**Fig. 26.** Comparison between GDR3 (top) and ZTF (bottom) light curves. The periodic source (bottom) is phase-folded to its detected timescale in each survey.

different  $Q$  index thresholds adopted in ZTF surveys, most QPS variables in Gaia were also QPS in ZTF. Some dippers in ZTF were identified as QPS or S variables in GDR3, although they are generally close to the  $M$  threshold in our study. Most dippers (QPDs or APDs) in Gaia were also dippers in ZTF. Most bursts in Gaia are detected as symmetric in ZTF, although we note that both surveys' low cadence makes them inefficient in detecting bursts. Fig. 26 shows two examples of sources in common in both surveys. GDR3 2067026678480142464 (top) shows similar behaviour in both surveys, and although we classified it as an L variable, its  $M$  index suggests the same bursting behaviour, with a slight shift towards higher symmetry in GDR3 than in ZTF. GDR3 216701033028221312 (bottom) has a similar behaviour in both surveys, with matching timescales but higher phase dispersion in ZTF than in GDR3, resulting in a higher  $Q$  index value for ZTF. Additionally, the ZTF light curve contains significantly more data points, contributing to its more symmetric behaviour compared to GDR3.

## 7. Summary and conclusions

We presented the results of a pilot study employing the asymmetry and periodicity variability metrics to characterise an all-sky sample of YSO candidates based on their G-band photometric variability observed with Gaia. We adapted the method behind the  $Q-M$  indices for application to Gaia's long-term but sparse light curves. Starting from the sample of 79 375 variables identified in GDR3 as YSO candidates based on their



**Fig. 27.** Q–M variation across surveys for sources observed in both GDR3 (black dots) and ZTF (colour dots). Line segments highlight Q–M variation across surveys, with colour matching their ZTF class.

variability, we pre-processed sources to filter for a low number of epochs, remove contaminants, and identify sources that were likely affected by spurious instrumental variability. Based on a sample of 28 179 sources, we inferred characteristic variability timescales by considering the possibility of periodic and aperiodic variability. The timescale distribution revealed a bimodality with rotational and intermediate timescales, with the former showing, on average, lower variability amplitudes than the latter. Rotational timescales were more frequent for sources with a higher periodicity and intermediate timescales for aperiodic sources. We successfully derived Q–M indices for 23 213 sources. We complemented our analysis with observations of  $H\alpha$  as a proxy of ongoing accretion and  $\alpha_{IR}$  as a proxy of the presence of circumstellar material and the evolution stage of YSOs. We found that WTTS discless sources tend to show more periodic and symmetric variability, while CTTS and WTTS disc-bearing sources tend to be more asymmetric and less periodic. We also examined the results as a function of the estimated spectral types for the sample and found that QPS variability is much more common among M type stars, while the incidence of L variability increases towards earlier types.

We used the derived Q–M indices to sort sources into one of seven variability morphology classes: Strictly periodic, quasi-periodic symmetric, stochastic, quasi-periodic dipper, aperiodic dipper, burst, or eclipsing binary. The timescales of 4966 additional sources were longer than half of the total observed time, and these were classified as long trend variables with only their M index derived. The subsamples in each of these

classes were further examined in terms of Q–M indices, variability amplitudes, timescales, and accretion and disc proxies, and we discussed possible links to well-known variability mechanisms. This examination revealed a wider variety of variability behaviour than previous studies, which we attribute to the longer time span of Gaia. We found sources displaying the usual variability caused by rotational modulation by magnetically induced spots or accretion-induced hot spots, and both bursting and dipping behaviour at rotational timescales, likely linked to star-disc interaction. Moreover, we found a relevant fraction of sources showing dipping and bursting behaviour at an intermediate timescale, which we tentatively linked to the phenomenology behind RW Aur fades and DQ Tau-type bursting variability.

We qualitatively compared our Q–M-based results with previous studies, including both high-cadence (CoRoT and K2) and ground-based surveys (ZTF), and discussed them at the individual and population levels. Although large survey-to-survey differences were found, the M index trends seem generally consistent. When considering survey-specific biases (caused by cadence, time span, precision, and age), we found that differences in the incidence of morphological classes across surveys can be explained. We note a lack of available observed datasets with common sources observed by Gaia and high-cadence surveys, which would allow us to probe the appropriateness of Q–M thresholds for a classification across surveys. Future studies should consider assessing the effect of survey-to-survey variations on these thresholds based on simulated data. Finally, Gaia DR4 will allow us to expand these results to an even broader all-sky coverage because the time span of the observations is longer and the calibration is improved.

## Data availability

The data table described in Table D.1 is available at the CDS via <https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/705/A172>.

*Acknowledgements.* We acknowledge funding from the European Union’s Horizon 2020 research and innovation program (grant agreement No. 101004141, NEMESIS). G.M. acknowledges support from the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We thank Berry Holl for helpful discussions on the use of Gaia DR3 data, Ann-Marie Cody and Laura Venuti for exchanges about the implementation of the Asymmetry and Periodicity metrics. This research was also supported by the International Space Science Institute (ISSI) in Bern, through ISSI International Team project 521 selected in 2021, Revisiting Star Formation in the Era of Big Data (<https://teams.issibern.ch/starformation/>) This work made use of Astropy (Astropy Collaboration 2013, 2018, 2022), Scikit-learn: Machine Learning in Python (Pedregosa et al. 2011), Pandas (pandas development team 2020), Numpy (Harris et al. 2020), and Matplotlib (Hunter 2007). This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

## References

- Alencar, S. H. P., Teixeira, P. S., Guimarães, M. M., et al. 2010, *A&A*, 519, A88
- Andre, P., Ward-Thompson, D., & Barsony, M. 1993, *ApJ*, 406, 122
- Andrae, R., Fouesneau, M., Sordo, R., et al. 2023, *A&A*, 674, A27
- Astropy Collaboration (Robitaille, T. P., et al.) 2013, *A&A*, 558, A33
- Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, *AJ*, 156, 123
- Astropy Collaboration (Price-Whelan, A. M., et al.) 2022, *ApJ*, 935, 167

- Auvergne, M., Bodin, P., Boisnard, L., et al. 2009, *A&A*, 506, 411
- Baluev, R. V. 2008, *MNRAS*, 385, 1279
- Barrado y Navascués, D., & Martín, E. L. 2003, *AJ*, 126, 2997
- Basri, G., Streichenberger, T., McWard, C., et al. 2022, *ApJ*, 924, 31
- Bouvier, J., Chelli, A., Allain, S., et al. 1999, *A&A*, 349, 619
- Bouvier, J., Grankin, K., Ellerbroek, L. E., Bouy, H., & Barrado, D. 2013, *A&A*, 557, A77
- Bredall, J. W., Shappee, B. J., Gaidos, E., et al. 2020, *MNRAS*, 496, 3257
- Buder, S., Kos, J., Wang, X. E., et al. 2025, *PASA*, 42, e051
- Cabrit, S., Pety, J., Pesenti, N., & Dougados, C. 2006, *A&A*, 452, 897
- Carpenter, J. M., Hillenbrand, L. A., & Skrutskie, M. F. 2001, *AJ*, 121, 3160
- Cieza, L., & Baliber, N. 2007, *ApJ*, 671, 605
- Cody, A. M., & Hillenbrand, L. A. 2018, *AJ*, 156, 71
- Cody, A. M., Stauffer, J., Baglin, A., et al. 2014, *AJ*, 147, 82
- Cody, A. M., Hillenbrand, L. A., David, T. J., et al. 2017, *ApJ*, 836, 41
- Cody, A. M., Hillenbrand, L. A., & Rebull, L. M. 2022, *AJ*, 163, 212
- Creevey, O. L., Sordo, R., Pailler, F., et al. 2023, *A&A*, 674, A26
- De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, *A&A*, 674, A2
- Delchambre, L., Bailer-Jones, C. A. L., Bellas-Velidis, I., et al. 2023, *A&A*, 674, A31
- Dembinski, H., Ongmongkolkul, P., Deil, C., et al. 2020, scikit-hep/iminuit (Zenodo), <https://doi.org/10.5281/zenodo.3949207>
- Distefano, E., Lanzafame, A. C., Brugaletta, E., et al. 2023, *A&A*, 674, A20
- Edenhofer, G., Zucker, C., Frank, P., et al. 2024, *A&A*, 685, A82
- Evans, D. W., Eyer, L., Busso, G., et al. 2023, *A&A*, 674, A4
- Eyer, L., & Genton, M. G. 1999, *A&AS*, 136, 421
- Eyer, L., Audard, M., Holl, B., et al. 2023, *A&A*, 674, A13
- Facchini, S., Manara, C. F., Schneider, P. C., et al. 2016, *A&A*, 596, A38
- Findeisen, K., Cody, A. M., & Hillenbrand, L. 2015, *ApJ*, 798, 89
- Fischer, W. J., Hillenbrand, L. A., Herczeg, G. J., et al. 2023, in *Astronomical Society of the Pacific Conference Series*, 534, Protostars and Planets VII, eds. S. Inutsuka, Y. Aikawa, T. Muto, K. Tomida, & M. Tamura, 355
- Gaia Collaboration (Prusti, T., et al.) 2016, *A&A*, 595, A1
- Gaia Collaboration (Bailer-Jones, C. A. L., et al.) 2023a, *A&A*, 674, A41
- Gaia Collaboration (Vallenari, A., et al.) 2023b, *A&A*, 674, A1
- Gaia ESA Archive. official webpage, <https://gea.esac.esa.int/archive/>
- Genova, F., Egret, D., Bienaymé, O., et al. 2000, *A&AS*, 143, 1
- Gilmore, G., Randich, S., Worley, C. C., et al. 2022, *A&A*, 666, A120
- Ginsburg, A., Sipińcz, B. M., Brasseur, C. E., et al. 2019, *AJ*, 157, 98
- Green, G. 2018, *J. Open Source Softw.*, 3, 695
- Greene, T. P., Wilking, B. A., Andre, P., Young, E. T., & Lada, C. J. 1994, *ApJ*, 434, 614
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Nature*, 585, 357
- Herbst, W., Herbst, D. K., Grossman, E. J., & Weinstein, D. 1994, *AJ*, 108, 1906
- Hillenbrand, L. A., Kiker, T. J., Gee, M., et al. 2022, *AJ*, 163, 263
- Holl, B., Fabricius, C., Portell, J., et al. 2023, *A&A*, 674, A25
- Hourihane, A., François, P., Worley, C. C., et al. 2023, *A&A*, 676, A129
- Hughes, P. A., Aller, H. D., & Aller, M. F. 1992, *ApJ*, 396, 469
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, 9, 90
- James, F., & Roos, M. 1975, *Comput. Phys. Commun.*, 10, 343
- Jiang, S. D., & Hillenbrand, L. A. 2024, *AJ*, 167, 221
- Joy, A. H. 1945, *ApJ*, 102, 168
- Kozłowski, S. 2016, *ApJ*, 826, 118
- Kulkarni, A. K., & Romanova, M. M. 2008, *MNRAS*, 386, 673
- Lada, C. J. 1987, in *Star Forming Regions*, 115, eds. M. Peimbert, & J. Jugaku, 1
- Lakeland, B., & Naylor, T. 2022, in *The 21st Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun*, Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun, 111
- Lomb, N. R. 1976, *Ap&SS*, 39, 447
- Marton, G., Abraham, P., Rimoldini, L., et al. 2023, *A&A*, 674, A21
- Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, *PASP*, 131, 018003
- McGinnis, P. T., Alencar, S. H. P., Guimarães, M. M., et al. 2015, *A&A*, 577, A11
- Morales-Calderón, M., Stauffer, J. R., Hillenbrand, L. A., et al. 2011, *ApJ*, 733, 50
- Mowlavi, N., Holl, B., Lecoeur-Taïbi, I., et al. 2023, *A&A*, 674, A16
- Natta, A., Grinin, V. P., Mannings, V., & Ungerechts, H. 1997, *ApJ*, 491, 885
- Ochsenbein, F., Bauer, P., & Marcout, J. 2000, *A&AS*, 143, 23
- Ordenes-Huanca, C., Zoccali, M., Bayo, A., et al. 2024, *MNRAS*, 533, 841
- pandas development team, T. 2020, pandas-dev/pandas: Pandas
- Pecaut, M. J., & Mamajek, E. E. 2013, *ApJS*, 208, 9
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, 12, 2825
- Petrov, P. P., Gahm, G. F., Djupvik, A. A., et al. 2015, *A&A*, 577, A73
- Pettersson, B., Armond, T., & Reipurth, B. 2014, *A&A*, 570, A30
- Randich, S., Gilmore, G., Magrini, L., et al. 2022, *A&A*, 666, A121
- Rimoldini, L., Holl, B., Gavras, P., et al. 2023, *A&A*, 674, A14
- Rodríguez, J. E., Pepper, J., Stassun, K. G., et al. 2013, *AJ*, 146, 112
- Roelens, M., Eyer, L., Mowlavi, N., et al. 2017, *MNRAS*, 472, 3230
- Romanova, M. M., Ustyugova, G. V., Koldoba, A. V., & Lovelace, R. V. E. 2013, *MNRAS*, 430, 699
- Roquette, J., Alencar, S. H. P., Bouvier, J., Guarcello, M. G., & Reipurth, B. 2020, *A&A*, 640, A128
- Roquette, J., Audard, M., Hernandez, D., et al. 2025, *A&A*, 702, A63
- Saunders, E. S., Naylor, T., & Allan, A. 2006, *Astron. Nachr.*, 327, 783
- Savitzky, A., & Golay, M. J. E. 1964, *Analyt. Chem.*, 36, 1627
- Scargle, J. D. 1982, *ApJ*, 263, 835
- Shulman, S. G., & Grinin, V. P. 2022, *MNRAS*, 512, 3098
- Simonetti, J. H., Cordes, J. M., & Heeschen, D. S. 1985, *ApJ*, 296, 46
- Sousa, A. P., Alencar, S. H. P., Bouvier, J., et al. 2016, *A&A*, 586, A47
- Stauffer, J., Cody, A. M., Baglin, A., et al. 2014, *AJ*, 147, 83
- Stauffer, J., Cody, A. M., Rebull, L., et al. 2016, *AJ*, 151, 60
- Stelzer, B., Flaccomio, E., Briggs, K., et al. 2007, *A&A*, 468, 463
- Tofflemire, B. M., Mathieu, R. D., Ardila, D. R., et al. 2017, *ApJ*, 835, 8
- VanderPlas, J. T. 2018, *ApJS*, 236, 16
- Venuti, L., Bouvier, J., Irwin, J., et al. 2015, *A&A*, 581, A66
- Venuti, L., Cody, A. M., Rebull, L. M., et al. 2021, *AJ*, 162, 101
- Wang, X.-L., Fang, M., Herczeg, G. J., et al. 2023, *Res. Astron. Astrophys.*, 23, 075015
- Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, *A&AS*, 143, 9
- White, R. J., & Basri, G. 2003, *ApJ*, 582, 1109
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, 140, 1868
- Zechmeister, M., & Kürster, M. 2009, *A&A*, 496, 577
- Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, *Res. Astron. Astrophys.*, 12, 723

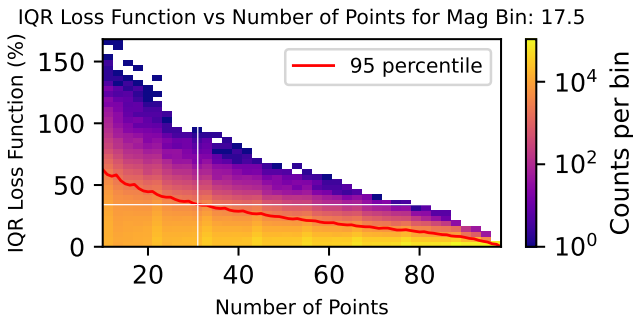
## Appendix A: Details on light curve filtering steps

### A.1. Bootstrap test for minimum number of epochs

We performed bootstrap tests to determine the minimum number of epochs required to ensure statistically robust variability indices. As a baseline, we selected light curves of GAPS sources that contained the highest number of epochs (typically `num_selected_g_fov` between 70 and 100 observations). To account for variations of signal-to-noise as a function of observed magnitude, we selected 5 sources per bin with 0.5 mag width for  $G$  between 6.5 and 20 mag. For each light curve selected, we computed the interquartile range (IQR) as a function of the number of epochs while randomly removing epochs until 10 epochs were left. To quantify the change introduced to the IQR by the removal of epochs, we computed a loss function as:

$$L_{IQR}(i) = \frac{|IQR_{original} - IQR(i)|}{IQR_{original}} \times 100, \quad (\text{A.1})$$

where  $IQR_{original}$  is the IQR of the original light curve and  $IQR(i)$  the IQR at the  $i^{th}$  step of epoch removal. The loss function represents a percentage difference between the original IQR and the IQR when points are removed. We repeated the process 10 000 times per light curve to minimise biases introduced by specific epochs while randomly removing epochs. An example of the distribution of loss function values as a function of the number of epochs is shown in Fig. A.1 for the mag bin  $G = 17.5$  mag from where we inferred that a minimum of 31 epochs was required to keep the relative changes in the IQR under  $1\sigma$  (34.1%) for 95% of the light curves in that magnitude bin (white cross).

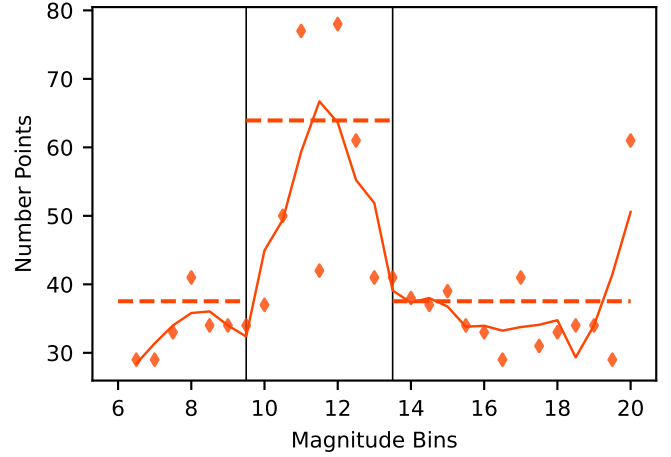


**Fig. A.1.** IQR variations as a function of the number of epochs for bootstrapped light curves in the magnitude bin 17.5 mag. The red line shows the 95<sup>th</sup> percentile of the distribution. The white cross indicates the threshold adopted for this magnitude bin.

Fig. A.2 shows the minimum number of epochs as a function of magnitude for all magnitude bins considered. There is a strong peak between 9.5 and 13.5 mags, meaning sources with this magnitude have an IQR changing more intensely, which is likely related to instrumental effects of Gaia around magnitude 12. To account for this feature, we applied a two-regime condition for the minimum number of observations and required at least 62 epochs for sources with magnitudes between 9.5 and 13.5 and 37 epochs for others.

### A.2. Spurious variability identification procedure

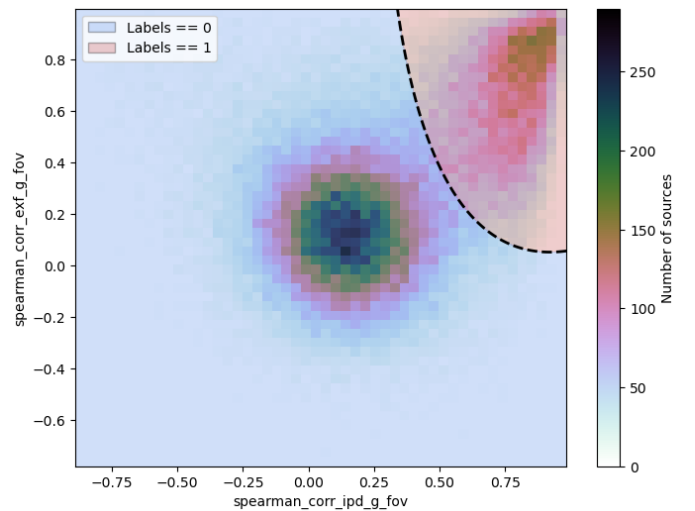
As part of GDR3, Holl et al. (2023) provides a series of parameters calculated to facilitate the identification of spurious



**Fig. A.2.** Minimum number of epochs needed to ensure the 95% of sources have an IQR variation under 34.1%, per mag bin. Diamonds show the raw values per mag bin. The solid line is a smoothed version. Vertical dotted lines show the 90 percentile of the points for a) mag bins between 9.5 and 13.5; b) other mag bins

variability signals introduced by the complex scanning law of Gaia and varying object scanning angles. These parameters were retrieved for our sources from the `vari_spurious_signals` table through [Gaia ESA Archive](#). In specific, we employed the pre-calculated Spearman correlation between photometric magnitudes and IPD GoF (Image Processing Data Goodness of Fit), `spearman_corr_ipd_g_fov`, and between magnitudes and their corrected flux excess factor `spearman_corr_exf_g_fov`. In both cases, a significant correlation suggests that scanning-law effects are likely present in the light curve.

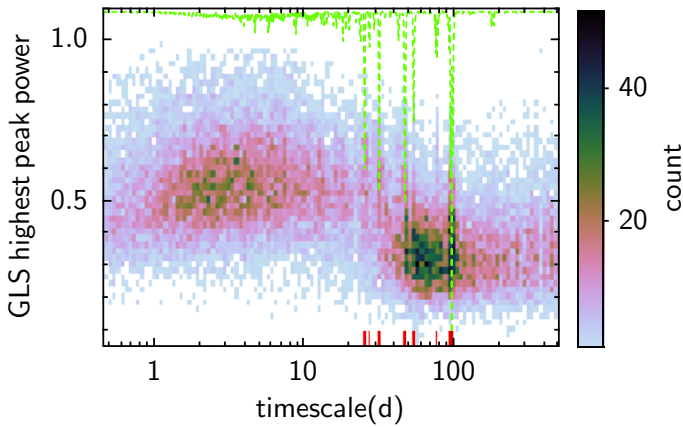
Fig. A.3 shows the distribution of `spearman_corr_exf_g_fov` vs `spearman_corr_ipd_g_fov`, where the distribution is seen split into two bulks. We used a Gaussian Mixture model to isolate these two bulks, which allowed us to flag 22 375 sources in the YSO sample as likely affected by spurious variability. 187 sources had no `spearman_corr_exf_g_fov` and were also discarded.



**Fig. A.3.** Distribution of Spearman correlations between the Gaia field-of-view photometric magnitudes and their corrected flux excess factor (y-axis) and their Image Processing Data Goodness of Fit (x-axis) for 79 188 GDR3 YSOs. The distribution was split into separate Gaussian distributions, delimited by the blue and red shading.

### A.3. Timescales in excess due to residual spurious variability

In Fig. A.4, we show a distribution of the power of the highest peak in the GLS periodogram as a function of the timescale for that peak (derived in Sect. 4.2.1). An excess of sources detected at specific long timescales (e.g.  $\sim 100$ , 50 and 25 d) is noticeable in this distribution. Fig. A.4 also shows a normalised frequency distribution of GLS timescales for sources affected by spurious variability identified in the previous section. Considering the timescales with at least 10% frequency among sources with spurious variability points to a correspondence of these with the timescales in excess, suggesting a prevalence of sources with spurious variability in our sample. We have further inspected these sources in the context of `spearman_corr_ipd_g_fov` and `spearman_corr_exf_g_fov` parameters discussed in the previous section to address if a different strategy could reduce their incidence. For example, the requirement that rather than both, either of these parameters show stronger correlation or using a lower correlation threshold. However, a relevant population of residual spurious variability sources was found even among sources with correlation values close to zero. In particular, we note that the recurrence of residual spurious variability is minimal within the primary (highly) variable sample defined in Sect. 3.5, but accentuated among sources in the secondary sample. Finally, we minimised this residual contamination by utilising the normalised frequency distribution of GLS timescales for sources affected by spurious variability identified in the previous section to mask the most frequent timescales among sources with spurious variability.

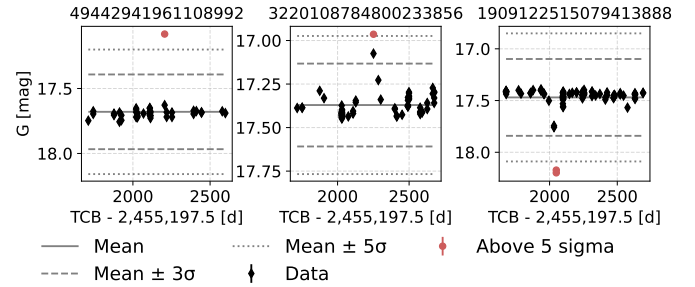


**Fig. A.4.** Density distribution of the power of the highest peak in the GLS periodograms as a function of their timescale for 34 739 variable sources analysed in Sect. 4.2.1, shown regardless of their FAL. The green dashed lines show the normalised frequency distribution of GLS timescales for sources affected by spurious variability introduced by the complex scanning law of Gaia and varying object scanning angles identified in Sect. A.2. This distribution is renormalised here for visualisation purposes. The red rugged lines highlight timescales in excess in our sample due to residual spurious variability, and masked from our sample.

### A.4. Procedure for outlier removal

Visual inspection of light curves exhibiting the highest peak-to-peak amplitude revealed that many sources had their variability and asymmetry analysis dominated by only one or two epochs beyond the  $5\sigma$  level. In the context of the Gaia sparse light curves, we must address our capacity to judge whether these are spurious outliers or the results of the sparse sampling of a

rapidly occurring variability event. For example, short-duration bursting features reported by Cody et al. (2017) typically span 10–20 days and typically show short duty cycles (time spent bursting). At certain latitudes, the Gaia scanning law can occasionally yield the occurrence of a dozen observed epochs over a couple of days, a phenomenon often referred to as cusps (see, for example, source *GDR3* 3336474807652948992 in Fig. 9). However, if not coincident with cusps, the sparse sampling of Gaia typically includes only one epoch every 10–20 days, henceforth under-resolving and, likely, often missing short-duration bursts.

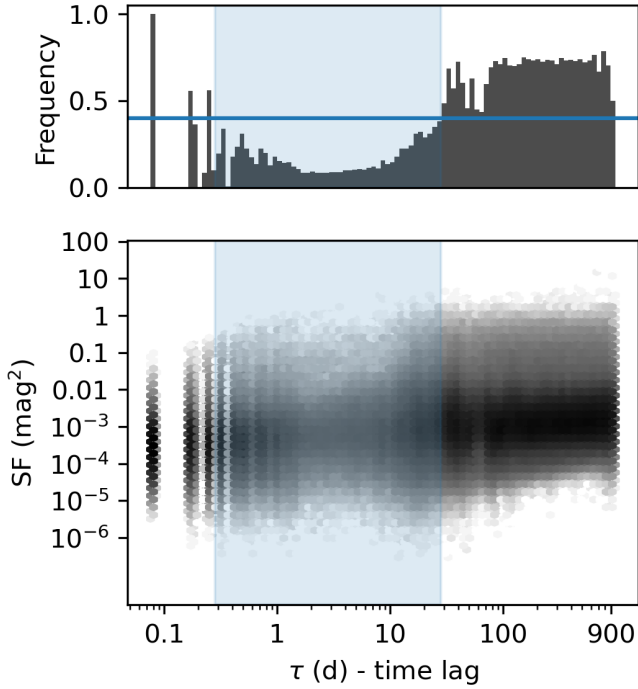


**Fig. A.5.** Examples of light curves with epochs beyond  $5\sigma$  level. *Left:* This epoch was not followed or preceded by a  $3\sigma$  outlier, and it was thus identified as an outlier and removed from the light curve. *Middle:* an example with a preceding epoch beyond the  $3\sigma$  level considered as a reliable variability feature. *Right:* Example with two successive epochs outside the  $5\sigma$  level also considered as a reliable variability feature.

We identified 735 sources (2% of the variable sample defined by Sect. 3.4) with at least one  $5\sigma$  outlier and only 11 having two outliers at the same level. We further hypothesise that finding these outliers at a systematically bursting behaviour could support the idea that these are undersampled but reliable rapid burst variables. However, we find an overall homogeneous distribution of outlier behaviour with about  $\sim 45\%$  flaring outliers versus  $\sim 55\%$  dimming outliers, thus denying our hypothesis. Including colour information could aid in further probing such events; nevertheless, colour variability was beyond the scope of the present study. We thus limit our approach to identifying and removing isolated outliers as follows. For each light curve with an epoch outside the  $5\sigma$  level, we evaluated if this epoch was isolated by examining if the immediately close epochs were at least within the  $3\sigma$  level while following the same bursting/dimming behaviour. If the epoch beyond the  $5\sigma$  level failed this criterion, it was considered an outlier and removed from the light curve. Fig. A.5 shows examples where the first light curve shows an example with an outlier, and two others show light curves with epochs outside the  $5\sigma$  level but identified as part of a reliable variability feature. 636 light curves had outliers removed with this criteria.

## Appendix B: Blind timescales

In this Appendix, we investigate the range of timescales affected by the sparse time coverage of Gaia. To this end, we employed Eq. (3) to calculate structure functions for all 43 840 sources with enough epochs in their light curves according to the procedure discussed in Sect. 3.2. Fig. B.1 shows the distribution of calculated  $\tau$ -SF( $\tau$ ) pairs at a population level, where a lower density of points is clearly seen at  $\tau$  values around  $\sim 1$ –10 days. The frequency of  $\tau$ -SF( $\tau$ ) pairs occurrence is also shown in the top panel of Fig. B.1. We choose a conservative requirement of at least 40% to define a blind-SF range of timescales in the range



**Fig. B.1.** *Bottom:* Structure function and time difference pairs considering all SF-points ( $\tau$ , SF) for all the 43 840 sources with enough number of epochs (Sect. 3.2). The blue region shows the depletion of pairs with  $\tau$  between 0.3 and 28 days, introduced by limitation in the Gaia time-series sampling. *Top:* Distribution of  $\tau$  as a function of the frequency in which sources had a valid SF( $\tau$ ). The blue horizontal line shows the 40% limit used to infer the blind timescale range of Gaia.

$\sim 0.3$ –28 days. Nevertheless, we note that the issue of underpopulated SF is more critical for the timescale range of  $\sim 0.5$ –10 days where any given  $\tau$ -SF( $\tau$ ) exists for less than 10% of sources. GDR3 scanning law (Fig. 1) implies that light curves will have varied sampling. Hence, the definition of this blind timescale range does not mean that timescales in this range cannot be measured with the SF method, but rather that SF will less frequently be well-sampled enough to detect such timescales. Based on the identification of this blind range, Sect. 3.2 includes additional processing steps to identify if an aperiodic timescale in this range came from an adequately sampled SF. Finally, we also note that this sampling issue is only expected to affect the SF method with the underlying assumption of an aperiodic SF model. For GLS periodic timescales, different underlying assumptions of a sinusoidal periodic signal apply, and various previous studies based on GDR3 light curves have corroborated its efficiency in detecting timescales in this range (e.g. Distefano et al. 2023).

## Appendix C: YSO properties

### C.1. Ongoing accretion

We used the data described in Sect. 2.2.2 to classify sources as CTTS or WTTS based on their  $H\alpha$  emission. For equivalent width measurements, we employed spectral type-dependent criteria by Barrado y Navascués & Martín (2003) converted into Gaia  $G_{BP}$ - $G_{RP}$  colours using the Pecaut & Mamajek (2013) empirical sequence. For full widths at 10% of the peak flux measurements, we used the White & Basri (2003) criterion to distinguish between CTTS and WTTS. For 64 sources, we used qualitative flags that report  $H\alpha$  in emission (from Pettersson

et al. 2014). When multiple  $H\alpha$  measurements were available, we considered that one measurement indicating a CTTS classification was enough to grant the source a classification as such. Specifically for EW( $H\alpha$ ) measurements from ESP-ELS Gaia DR3 we have restricted our analysis to 7318 sources with the absolute value of reported EW( $H\alpha$ ) larger than at least 5 times the reported uncertainties. This decision was motivated by the limitations due to  $G_{BP}/G_{RP}$  spectral low resolution ( $R \sim 30$ –100; De Angeli et al. 2023) and its representation as Hermite function bases, where the identification of sources with large values of EW( $H\alpha$ ) are reliable, but where low accreting and non-accreting sources with EW( $H\alpha$ ) values closer to the stellar continuum may be confused within uncertainties.

### C.2. Infrared indices

To estimate  $\alpha_{IR}$  we followed a similar procedure as the one in the NEMESIS YSO catalogue (Roquette et al. 2025): We used the Vizier Photometry viewer tool to retrieve for each source all previously published photometric data available at the Centre de données astronomiques de Strasbourg (Genova et al. 2000; Ochsenbein et al. 2000), and used all data available for the wavelength range 2–24  $\mu\text{m}$  to fit a linear model and measure  $\alpha_{IR}$ , with a requirement of at least two photometric data points 2  $\mu\text{m}$  apart in this interval. Next, we adapted the standard classification scheme for YSOs based on their  $\alpha_{IR}$  (Lada 1987; Andre et al. 1993; Greene et al. 1994) by grouping sources based on the dominant contributor to their SED: sources with significant envelope contribution have  $\alpha_{IR} > -0.3$  (1378 sources), sources with a thick disc have  $-0.3 < \alpha_{IR} \leq -1.6$  (6427 sources), sources with thin discs have  $-1.6 < \alpha_{IR} \leq -2.5$  (8743 sources) and discless sources have  $\alpha_{IR} < -2.5$ , (11 631 sources).

### C.3. Spectral types

In GDR3, effective temperatures from the general stellar parameterizer from photometry (GSP-Phot; Andrae et al. 2023) were available for 21 112 of our sources. We combined these with the  $T_{\text{eff}}$ -Spectral Type scale for PMS stars by Pecaut & Mamajek (2013) to infer spectral types for our sample. For 7067 sources, effective temperatures were unavailable, and instead we used a  $(G_{BP} - G_{RP})_0$ -Spectral Type version of the Pecaut & Mamajek (2013) scale. For this end, individual extinctions were inferred at the position and the GDR3 distance of sources using the Galactic 3D dust map by Edenhofer et al. (2024) with the dustmaps Python package (Green 2018). Edenhofer et al. (2024) provides a robust map for sources up to 1.25 kpc, with a secondary map based on less data available up to 2 kpc. For 84 sources located at larger distances, we adopted the 2 kpc extinction value as a bottom limit. Although informative, stellar parameters for young stars provided by GSP-Phot have a series of caveats (see Andrae et al. 2023, for an in-depth discussion), including the use of time-averaged mean BP/RP spectra, which neglects YSO variability, and known degeneracy between effective temperature and extinction - expected to be especially relevant for sources in regions of enhanced line-of-sight extinction such as star-forming regions. Due to these known limitations, we did not attempt to derive sub-spectral types, and we stress that the reported spectral types are rough estimations aimed at a broad insight into the content of our YSO sample.

**Table D.1.** Overview of the table contents

Name	Description	Unit
source_id	GDR3 source identifier	
ra	Right ascension in ICRS at Gaia reference epoch	degree
dec	Declination in ICRS at Gaia reference epoch	degree
parallax	Absolute stellar parallax at Gaia reference epoch	mas
phot_g_mean_mag	G-band mean magnitude	mag
num_selected_g_fov	Number of G FOV transits selected for variability analysis by Gaia	
CTTS	If 0, H $\alpha$ observation exist but does not indicate ongoing accretion (WTTS). If > 0, at least one observation reports ongoing accretion (CTTS). If null, no available information. See Appendix C.1.	
alpha_ir	$\alpha_{IR}$ index. See Appendix C.2.	
disc <sup>a</sup>	Infrared classification as in Appendix C.2	
nsdpoints	number of points in the SED available for $\alpha_{IR}$ derivation	
min_wl	minimum wavelength available for $\alpha_{IR}$ derivation	$\mu\text{m}$
max_wl	maximum wavelength available for $\alpha_{IR}$ derivation	$\mu\text{m}$
timescale_method	GLS if the GLS periodogram was used to get the timescale (Sect. 4.2.1). SF if the structure function was used (Sect. 4.2.2).	
timescale	Dominant timescale, if not larger than half the light curve duration	days
long_timescale_method	GLS if the GLS periodogram was used to get the timescale. SF if the structure function was used. Null if the dominant timescale is not long.	
long_timescale	Dominant timescale, if larger than half the light curve duration	days
primary	True if the source is part of the primary sample. False if the source is part of the secondary sample. See Sect. 3.5.	
Q_index	Q index value from Eq. 1	
M_index	M index value from Eq. 2	
amplitude	Peak-to-peak amplitude of the G-band light curve	mag
morphological_class	Morphological class inferred from Q–M indices. Can be P, EB, QPS, QPD, B, S, APD, L (Sect. 4.4).	
SpT	Spectral type estimation (App.C.3)	
SpT_method <sup>b</sup>	Method for Spectral type estimation	

**Notes.** <sup>(a)</sup>Discless: sources without disc contribution detectable in their SED ( $\alpha_{IR} \leq -2.5$ ); **Thin**: sources with detectable contribution of a thin disc ( $-2.5 \leq \alpha_{IR} \leq -1.6$ ); **Thick**: sources with detectable contribution of a thick disc ( $-1.6 \leq \alpha_{IR} \leq -0.3$ ) **Envelope**: sources with detectable contribution an envelope ( $\alpha_{IR} \geq -0.3$ ); <sup>(b)</sup>Te<sub>eff</sub>->SpT: Spectral Type estimated based on the comparison of the source GDR3 GSP-Phot temperature with an empirical sequence by Pecaut & Mamajek (2013) ; Bp-Rp->SpT Spectral type estimated based comparison of its  $(G_{BP} - G_{RP})_0$  with the empirical sequence.

## Appendix D: Description of the data table

Table D.1 provides an overview of the parameters used in this study, including source identifiers, astrometric and photometric properties, variability metrics, and classification results. The corresponding data table is available on Vizier.